Are Public Intrusion Datasets Fit for Purpose

Characterising the State of the Art in Intrusion Event Datasets

Kenyon A.
R&D, Hyperscalar Ltd, Guildford, United Kingdom, tony.kenyon@ieee.com

Deka L.
Institute of Artificial Intelligence, De Montfort University, Leicester, United Kingdom, lipika.deka@dmu.ac.uk

Elizondo D.
Institute of Artificial Intelligence, De Montfort University, Leicester, United Kingdom, elizondo@dmu.ac.uk

ABSTRACT

In recent years cybersecurity attacks have caused major disruption and information loss for online organisations, with high profile incidents in the news. One of the key challenges in advancing the state of the art in intrusion detection is the lack of representative datasets. These datasets typically contain millions of time-ordered events (e.g. network packet traces, flow summaries, log entries); subsequently analysed to identify abnormal behavior and specific attacks [1]. Generating realistic datasets has historically required expensive networked assets, specialised traffic generators, and considerable design preparation. Even with advances in virtualisation it remains challenging to create and maintain a representative environment.

Major improvements are needed in the design, quality and availability of datasets, to assist researchers in developing advanced detection techniques. With the emergence of new technology paradigms, such as intelligent transport and autonomous vehicles, it is also likely that new classes of threat will emerge [2]. Given the rate of change in threat behavior [3] datasets become quickly obsolete, and some of the most widely cited datasets date back over two decades. Older datasets have limited value: often heavily filtered and anonymised, with unrealistic event distributions, and opaque design methodology.

The relative scarcity of (Intrusion Detection System) IDS datasets is compounded by the lack of a central registry, and inconsistent information on provenance. Researchers may also find it hard to locate datasets or understand their relative merits. In addition, many datasets rely on simulation, originating from academic or government institutions. The publication process itself often creates conflicts, with the need to de-identify sensitive information in order to meet regulations such as General Data Protection Act (GDPR) [4]. Another final issue for researchers is the lack of standardised metrics with which to compare dataset quality.

In this paper we attempt to classify the most widely used public intrusion datasets, providing references to archives and associated literature. We illustrate their relative utility and scope, highlighting the threat composition, formats, special features, and associated limitations. We identify best practice in dataset design, and describe potential pitfalls of designing anomaly detection techniques based on data that may be either inappropriate, or compromised due to unrealistic threat coverage. Such contributions as made in this paper is expected to facilitate continuous research and development for effectively combating the constantly evolving cyber threat landscape.

CCS CONCEPTS

• Intrusion Detection • Intrusion Prevention • Anomaly Detection • Network Flow • Smart Cities

KEYWORDS

Intrusion dataset, intrusion detection, anomaly detection, intrusion prevention, ddos, malware, netfow, masquerade, nids

1 Introduction

Problem Statement: A key challenge for researchers working in intrusion detection today is the limited availability of high quality publically available datasets. We find that datasets are often hard to locate and frequently suffer from a range of issues, including:

- **Poor provenance** - unclear methodology on how the dataset was created
- **Unclear composition** - with respect to benign and malicious event frequencies and scope
- **Poor data quality** - such as duplicate data, unrepresentative samples
- **Over-summarisation**, - with potential loss of features, and inability to scrutinise origin data
- **Unmaintained 'frozen' data** -leading to unrepresentative threat event signatures
- **Limited to specific domains** - unrepresentative of commercial deployments (e.g. academic networks)
- **Under-represented domains** - such as industrial networks, IoT networks

The importance of high quality datasets: Cyber threats are becoming ever more sophisticated, and yet typically represent only a tiny fraction of overall event traffic, making accurate real-time detection an extremely hard problem, with only incremental gains in recent years. Increasingly we see machine learning techniques being used to detect such anomalies, across many domains, based on large representative datasets. The lack of representation and variable quality in intrusion detections means that detection models can be significantly compromised, if based on poor quality data.

Why the scarcity of good datasets: The cybersecurity domain introduces a number of unique challenges with regard to the creation of representative event datasets. Experimentation and data collection on live networks is normally infeasible, especially where that these systems have business- or mission-critical functions, or where the data contains sensitive information. Creating representative environments requires skillful design, potentially with large numbers of network assets (real or simulated servers, clients, routers, switches etc.). Specialised malware generators may be required, with well-insulated network infrastructure to limit damage from any malware or threat simulation.

All of these factors mean that IDS datasets are a serious challenge to create, and maintain, and those datasets that do exist may represent only a synthetic subset of infrastructure and the full range of threat event types. The scarcity of good public datasets severely impairs experimentation and evaluation of IDSs, particularly anomaly-based detectors [5]: [6] states: "the most significant challenge an evaluation faces is the lack of appropriate public datasets for assessing anomaly detection systems." Furthermore, there has been little research on the use of consistent metrics for measuring how realistic the data is with regard to live network and system performance, and the presence and distributions of anomalies and threat patterns.

State of the art: We find that many of the available public datasets for IDS research are static (i.e. they can be thought of as snapshots in time. The particular challenges around maintainability mean that authors often create datasets for specific research initiatives, after which motivation to provide updates diminishes - with a few notable exceptions, which we discuss later. Cybersecurity is a highly dynamic field; with threat behavior constantly evolving [7], Trend reports published by antivirus companies show that the number of unique malicious executable files has risen from less than one million to over one billion between 2008 and 2014 [8, 9]. Security companies that analyse malware now routinely collecting over one million unique files per day [3]. Threat sophistication and anomaly distributions also vary significantly by context; for example certain industries may be highly attractive as targets for hackers, and may therefore exhibit higher frequencies or specific types of malware. Later in this paper we illustrate how under-represented these datasets are for specific domains.

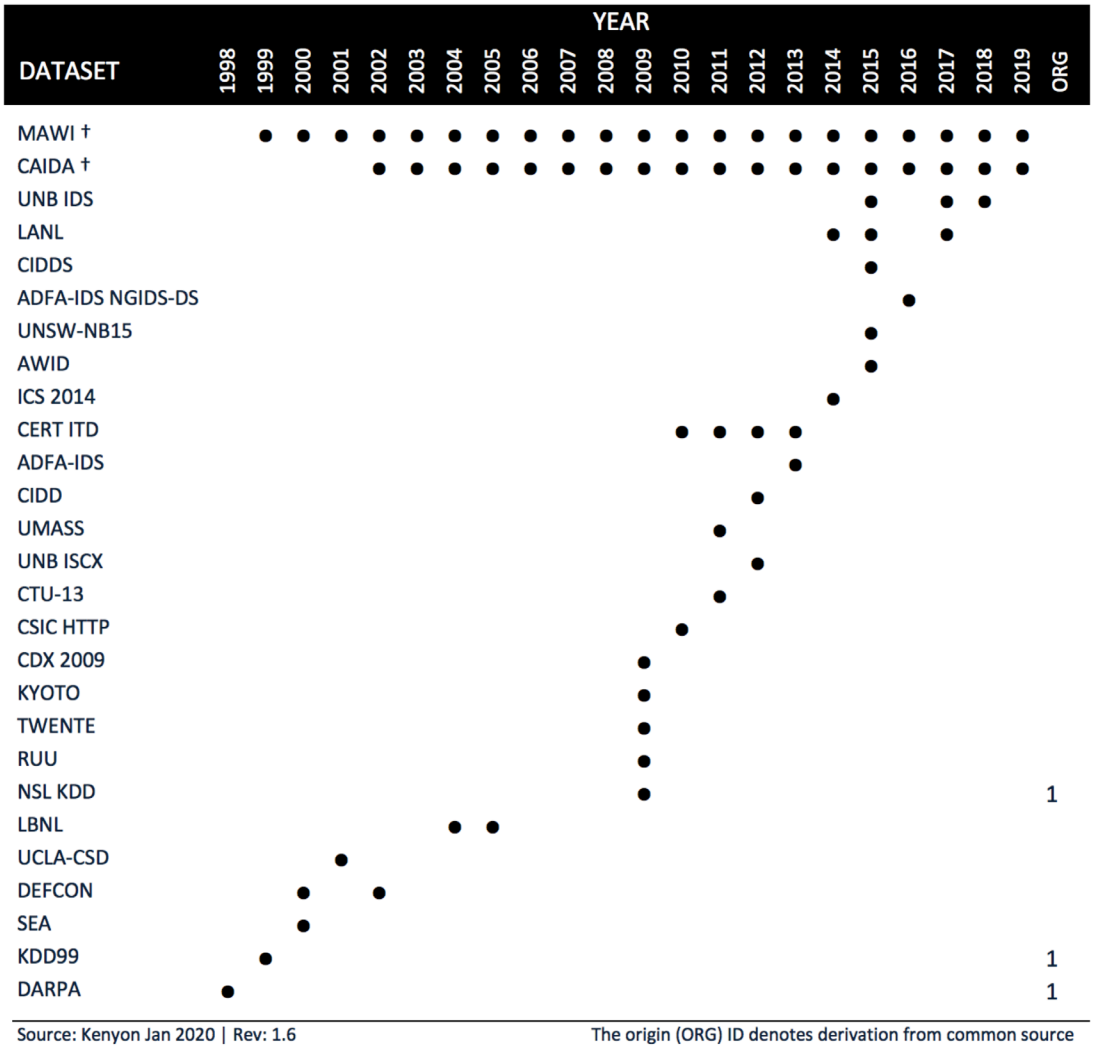


Figure 1: Major public intrusion dataset releases over time. Here we can clearly see the timeline of each dataset, as well as an indication of how well-maintained a dataset is over time. † Note all datasets are equivalent, the MAWI and CAIDA datasets for example are long-term archives of Internet backbone data. Section 3 provides detail on the composition of each dataset and the deployment context.

As a consequence, intrusion datasets have a tendency to become 'stale', especially if not regularly maintained. Figure 1 illustrates how many of the public intrusion datasets are unmaintained, and effectively obsolete (despite still being in active use today). For long-term consistent experimentation with high-quality data, alternative solutions need to be found, including broader cooperation and redacted publication of live representative data, and open simulation frameworks - where researchers can script scenarios and update threat vectors, and increased use of automated programmable cloud model creation.

Contribution: In this paper we uniquely analyse the broad corpus of public intrusion datasets with consistent scrutiny, up-to-date links to archives, references on important analysis, and transparency on provenance. We rank datasets according to relative utility, and offer high-level analysis on composition, methodology, maintenance, data quality issues, and relevance. We provide insights on how researchers should assess each dataset for accuracy, scope and currency (with respect to recent threat trends and the deployment context).

Finally, we identify best practice in dataset design, and describe potential pitfalls of designing anomaly detection techniques based on data that may be either inappropriate, or compromised due to unrealistic threat coverage. We anticipate that there will be three main beneficiaries of this paper:

- **Dataset designers** – 'best-practice' guidance and transparency on common errors should encourage designers not to repeat the mistakes of earlier dataset design, and help extend the life of datasets.
- **Security researchers** – are able to identify datasets that are most appropriate for their research problem, mindful of provenance, assumptions and compromises in the dataset design.
- **Industry-specific security researchers** - are able to identify deployment-specific datasets (e.g. backbone, cloud, IoT) most appropriate for their research problem. Where matching datasets are not available, researchers can make informed choices on the use of alternates.

There is very little consistent information in the literature with which to critically assess the full range of public intrusion datasets, since much of this material is fragmented, often as a partial critique when a new dataset is created. Recent work in this area is beginning to address this gap more formally [100],[101]. We provide citations throughout the text to reviews on specific datasets.

Some of the key issues important in dataset evaluation and selection are described in Section 2. In Section 3 we review examples of the most widely used public datasets available today. Section 4 analyses key flaws in dataset design and their implications for researchers, and offers guidance on best practice. Section 5 provides a concrete example what researchers might expect from a higher quality modern dataset.

2 Key Concepts in Intrusion Dataset Design

In this section we introduce important topics relevant to design, maintenance, and evaluation of datasets, including: classes of threat routinely present; tools commonly cited in dataset creation and analysis; timeliness and accuracy in detection models; how dataset design should reflect real-world threat scenarios; potential impacts of de-identification; techniques to validate dataset quality.

2.1 Threat Characterisation

Before we go further into describing the key concepts of intrusion datasets, it is important to quickly review the broad taxonomy of cybersecurity threats that we might expect to make up a modern IDS dataset, and the associated tools and nomenclature.

Cybersecurity intrusion comprises a number of different contexts and attack methodologies, the most common are summarized below. The first distinction we should note is the location of the attacker with respect to trust boundaries, geo-location, and privileged access to sensitive internal systems and information:

- **Insider threat** – the attacker resides within an organisation's security perimeter (the main trust boundary between public and private network infrastructure). For example, a disgruntled employee may have legitimate access to sensitive company assets, possibly with elevated privileges. Detecting insider threat is therefore particularly challenging since layered security controls may already be bypassed, and a great deal of instrumentation may be absent or hard to interpret when attempting to correlate events.
- **External threat** – the attacker resides outside an organisation's security perimeter, and must penetrate layered security controls. Skilled attackers may subvert controls by compromising insiders (through social engineering, blackmail etc.) or by installing malware - such as rootkits - on vulnerable machines within the security perimeter.

The second distinction we should note is the method of attack, resources employed, and key objectives. Note that the specific exploits here are evolving regularly, and for a dataset to hold value it should include both recent and legacy examples of the major attack classes below:

- **Denial of Service (DoS)** – attempts to render a service unavailable, either by flooding the network with high volume event traffic (a volumetric attack such as TCP SYNflood) or by gradually using up all the resources of the service (through resource starvation). Volumetric attacks are often easier to detect since they are easily detected and classified, resource starvation attacks may be much stealthier and use legitimate flow features that are much harder to classify.
- **Distributed Denial of Service (DDoS)** – a DDoS attack is essentially a DoS attack that is mounted from multiple networked systems (often from a compromised Botnet). Such attacks typically generate huge volumes of traffic with the potential to take down corporate networks, gaming servers, or even substantial parts of the Internet.
- **Botnet** – a botnet is a cluster of vulnerable networked devices that have been compromised by an attacker, and their resources and/or geo-location used by to perform various malicious actions; such as mounting a DDoS attack, stealing sensitive content, transmitting spam, Man-in-the-Middle (MiM) attacks or eavesdropping on sensitive traffic or video feeds.
- **Advanced Persistent Threat (APT)** – is a set of stealthy and continuous computer hacking processes, often orchestrated by well-organised individuals (possibly even state sponsored) targeting a specific organisation or entity. An APT often targets private organisations, but the scope can extend to entire states, usually for business or political motives. APTs usually employ a high degree of covertness over a long period of time.
- **Brute Force Attack** – a popular method for discovering authentication credentials and hidden content or web pages on servers. Brute force attacks normally work by iterating through many guesses, parsing web content for links, attempting to 'walk' a web tree etc.
- **Web Service Attacks** – many organisations today rely almost entirely on the Internet, and the technologies used to create web commerce sites are evolving rapidly, making them vulnerable to a broad range of attacks. The objectives of such attacks can be wide ranging: from stealing sensitive content, breaching Personally Identifiable Information (PII), or defacing a service to damage brand reputation. Attack techniques range from compromising vulnerabilities (e.g. SQL Injection, Cross-Site Scripting (XSS) etc.), to brute force attacks.
- **Infiltration Attack** – attackers can gain privileged access to sensitive internal assets through a range of malware techniques. Often these attacks are preceded by reconnaissance activities such as a port scan, where vulnerable TCP and UDP ports may be discovered using tools such as Network mapper (NMAP). Vulnerabilities in software can be exploited to elevate access privileges; with Rootkit malware installed to provide a 'backdoor' into a vulnerable machine. Once an attacker has privileged access they may spend months scanning the network for sensitive assets and executing further malicious acts.
- **Masquerade Attack** – is an attack that uses forged identities to gain unauthorised access to networks and systems. Attacks are often performed using stolen user credentials (e.g. via social engineering), by exploiting vulnerabilities in software or protocols, or by bypassing security controls entirely (e.g. by accessing an unlocked computer). The main characteristic of such an attack is that the attacker is masquerading as a legitimate user, and as such it may be very difficult to detect.
- **Malware** – malicious software designed specifically to harm a user or organisations assets or data. Malware comes in different forms, each with distinct features and behaviour: i) *Backdoor*, enables unauthorised access to compromised computers, ii) *Exploit*, uses a software vulnerability to gain authorized access, iii) *Virus*, is a self-replicating or parasitic infector, iv) *Worm*, a self-replicating stand-alone malware, v) *Trojan*, non-replicating software with hidden functionality. vi) *RootKit*, stealthy software that actively hides itself, vii) *SpyWare*, software that invades user privacy through information gathering, viii) *HackTool*, exploit, attack and scanning & reconnaissance tools and libraries.
- **Ransomware** - a type of malware that threatens to publish a victim's data, destroy or permanently block access to it, unless a ransom is paid. Advanced malware techniques such as 'cryptoviral extortion' may be employed to encrypt a victim's files, requiring a payment (often using an untraceable digital currency) to decrypt them.

Importantly, we should recognise that certain classes of threat are targeted at specific protocols and services, and that features important in the detection of one class of threat may be irrelevant to another. In developing detection models we should therefore be clear whether the objective is to identify broad classes of threat, or to detect specific threats more effectively. With this in mind we should consider whether the features available in datasets enhance or preclude developing such models. Several recent datasets described in

Section 3 include useful flow summaries and pre-prepared statistical features, as well as detailed event traces. Earlier datasets are often deficient in this regard, including only summary data, lacking in important timing information. Heavy de-identification may obscure key features. The lack of raw trace files will impair our ability to reconstruct flow summaries, the lack of timestamps and geo-location contextual data could preclude certain threat models. Naïve simulation in the data design may negate the validity of a dataset entirely.

There have been several proposed anomaly taxonomies. Mirkovic et al. [10] proposed a classification of DDoS attacks and defense mechanisms according to criteria such as IP address spoofing and exploited weakness. Barnett et al. [11] present a taxonomy of scanning events. Plonka et al. [12] offers a taxonomy covering a broad range of anomalies. CAPEC [13] provides an online database of host attack patterns. Mazel et al [14] offers a taxonomy of anomalies in wide are backbone traffic. Grégio et al [15] provides brief survey on available malware taxonomies. For further information on threat taxonomies and classifications refer to [16]. For information on recent trends refer to [7].

2.2 Representative Event Data

When choosing a dataset we need to be mindful of how representative the event data is: in terms of the frequencies of threat and normal events, the relative proportion of services and protocols included, whether the data has been recorded from live or simulated feeds, and the domain in which the events were generated. Obtaining representative, accurate, useful well-labeled training material is notoriously difficult, and maintaining these datasets often impractical. Paradoxically many organisations that have the scale and capability to publish useful data - and would benefit the significantly from advances in research - are highly protective of such information; not least because publishing has the potential to expose infrastructure vulnerabilities, as well as sensitive PII. The effort to anonymise such data is often viewed as prohibitive or an unacceptable risk.

In Figure 2 we list a number of important public intrusion datasets that have been widely used for modeling and analysing cybersecurity threats, and classifying anomalous behavior. We highlight information on the composition of the archive and its size. The repositories are listed by date, most recent updates first, since time is an important factor in determining the utility of a dataset (given the rapidly evolving nature of cybersecurity threats). Additional features are indicated, such as whether the events are real or simulated, or a hybrid; the deployment context; whether the events have been de-identified; and whether flow summaries are included. In Figure 3 we illustrate threat coverage by each dataset, and this gives a second important dimension to dataset utility. In cybersecurity we need always to be looking at both recent attack trends, as well as legacy threats. It should be noted that whilst several of these datasets are used widely in research and teaching today, they are effectively obsolete, for reasons we discuss in the following sections.

NAME	YEAR	ORIGIN	TYPE	FEATURES				FORMATS					SIZE
				ORG	ANO	LAB	FLW	PCP	TCP	CSV	XML	TXT	GB
MAWI	1999-2019	MAWI Working Group	Internet backbone. TCPdump. PCAP. Summaries	R	■	■	-	■	-	■	■	-	50+
CAIDA	2002-2019	Center of Applied Internet Data	Campus and internet backbone traffic	R	■	-	-	■	■	-	-	-	100+
CSE-CIC-IDS2018	2018	University of New Brunswick CSC & CIC	Enterprise network & service traffic. CSV flows, PCAP and logs	S	-	■	■	-	■	-	■	-	220+
CICIDS2017	2017	University of New Brunswick	Enterprise network & service traffic. CSV flows, PCAP	S	-	■	■	■	-	■	-	■	60
CIDDS	2017	Coburg University, Germany	Cloud network & service traffic (SME). NetFlow. External threats.	H	■	■	■	-	-	-	-	-	0.5
LANL	2014-2017	Los Alamos National Laboratory	Security and authentication cyber threats. TXT logs	R	■	■	■	-	-	■	-	-	100+
UNSW-NB15	2015	University of New South Wales	PCAP. BRO. Argus. CSV. Summaries	S	-	■	■	-	■	■	-	-	100
AWID	2015	University of the Aegean	802.11 WEP WiFi events	H	-	■	-	-	■	-	-	-	-
CERT ITTD	2010-2013	SEI at Carneigie Mellon University	Over 700 insider threats, authentications. CSV logs.	S	-	■	-	-	-	■	-	-	90+
ICS 2014	2014	ICS	Contains 28 threats on SCADA. Snort, WEKA, ARFF logs	S	-	■	-	-	-	■	-	■	0.1
ADFA-IDS	2016 2013	University of New South Wales	Database & web services cyber range. Host and network logs. CSV.	S	-	■	■	■	-	■	-	■	20
UNSW-NB15	2015	University of New South Wales	Database & web services cyber range. Host and network logs. CSV.	S	-	■	■	-	■	■	-	■	20
CIDD	2012	CIDD	Cloud masquerade data. XLS Logs	R	-	■	-	-	■	■	-	-	0.1
UMASS	2011	University of Massachusetts	Web based attack traces over wireless nets	S	-	■	-	■	■	-	-	-	-
ISCXIDS2012	2012	University of New Brunswick	Network and service traffic. XML flows, PCAP	S	-	■	■	■	-	-	■	-	35
CTU-13	2011	Stratosphere IPS	Live Botnet attacks and benign traffic (Spam, DDoS, Clickfraud etc).	R	-	■	■	■	-	-	-	■	2
CSIC HTTP	2010	CSIC Spanish National Research	WAF attack dataset. SQL injection, buffer overflows etc. CSV logs	H	-	■	-	-	-	■	-	-	0.1
CDX 2009	2009	UCMA	UCMA Network & Server logs. PCAP. Snort, Splunk. Dns, Apache Logs	R	-	-	-	■	■	-	-	■	13+
KYOTO	2009	Kyoto University	Honeypot traffic	S	-	-	-	-	-	-	-	-	-
TWENTE	2009	University of Twente	Web services. OpenSSH. Apache. Proftp. Netflow traces	S	-	-	■	-	-	-	-	-	-
NSL-KDD	2009	Tavallaee	Network Traffic. TCPDump. CSV summaries	S	-	■	-	-	-	■	-	-	0.1
RUU	2009	Columbia University IDS Group	Masquerade traffic, university campus. 48 users	S	■	-	-	-	-	-	-	-	na
LBNL	2004-2005	Lawrence Berkeley National Laboratory	Packet header traces from LBNL enterprise network (no payload)	?	■	■	-	-	■	-	-	-	-
DEFCON	2000-2002	The Shmoo Group	Port scans, bad packets, privilege escalation	S	-	-	-	-	-	-	-	-	-
UCLA-CSD	2001	UCLA	Border router attack traces. TXT logs	S	■	■	-	-	-	-	-	■	0.5
SEA	2000	Schonberg (PhD study data)	User Masquerade Behaviour, 50 users. Injected session behaviour.	S	■	-	-	-	-	-	-	■	0.01
KDD99	1999	University of California, Irvine	Network Traffic. TCPDump. CSV summaries	S	-	■	-	-	-	■	-	-	0.1
DARPA 98-99	1998	Lincoln Laboratory	Network Traffic. TCPDump binaries. CSV summaries	S	-	■	-	-	■	■	-	-	4

Source: Kenyon Jan 2020 | Rev: 3.3

Figure 2: Summary of major public intrusion detection event datasets, including a breakdown of key features, data formats and archive size. Features include: the dataset origin (ORG), where R indicates real data sources, and S indicates simulated data sources, and H indicates a hybrid; whether anonymisation (ANO) has been performed on the data, whether the data is labelled (LAB), and contains flow summaries (FLW). Formats include: PCAP formatted events, CSV, XML, or TXT.

It is important to note that the relative frequency of attack to benign events in real networks is usually extremely low, especially where event data is collected in bulk over long time periods (discussed further in Section 5). This is an important consideration when designing anomaly detection systems for use in real-world deployments, and important in dataset design. In the public datasets cited in Figure 2, many are synthesised, and a number have been generated in somewhat artificial or specialised contexts (e.g. military cyber-range simulations, academic networks, and Internet backbones): as such the event distributions and threat content may be heavily distorted and not represent real-world corporate and public networks. Some of the more recent datasets exhibit fairly realistic and general-purpose distributions (by realistic we mean closely corresponding with real world normal and abnormal event distributions, comparing a number of key metrics [18, 19]). In Section 3 we note such distortions in the dataset descriptions.

In recent years, dataset generation has become more sophisticated, and some of the newer datasets provide valuable insight into threat modeling and discovery, including extensive flow and summary metadata to support further analysis (such as UNB ISCX). The use of each dataset needs to be carefully considered depending on the requirement; traffic distribution for example is very sensitive to the deployment context. Further, each dataset represents a 'snapshot' of the threat landscape, often deliberately constrained in scope by the resources available to the publisher. Attack scenarios may be heavily manipulated or crudely synthesised. With older datasets there is the possibility of some threats not being labeled correctly, or simply not being present. The impact of de-identification can vary considerably. These subtleties may not be immediately transparent to researchers.

2.3 Dataset Quality and Transparency

There are clear challenges in obtaining, maintaining and publishing good quality intrusion datasets; there are always compromises. At the present time there are no consistent metrics on the quality a dataset, and researchers are often left to assess dataset quality based on the reputation of the authors, the quality of published results, and anecdotal evidence. In order to base further research on a dataset we must be able to verify that the dataset bears some resemblance to reality. In Section 4 we summarise the major defects in publics IDS datasets, and define some of the key attributes that we should look for in assessing data quality.

In more recently published datasets there have been significant attempts to ensure that generated data closely match real-world event characteristics. For example researchers at the University of New Brunswick perform extensive analysis of real packet traces in order to create profiles for traffic-generating agents, and publish guidelines for how to obtain valid datasets [20].

To date there has been little research on quantitative analysis on such datasets and no systematic metric for assessing quality. We need standard metrics for describing how realistic the data is with regard to true network and system performance, as well as the distributions of anomalies and threat patterns. This is a difficult task given the wide variability in deployment contexts and temporal traffic patterns across various organisations, however there are basic features we can begin to correlate in a quantifiable manner. Researchers at the University of New South Wales [19] propose a quality metric based on the Sugeno fuzzy inference model, and have produced a synthetic intrusion dataset (ADFA NG-IDS) that correlates strongly with realistic data, using commercial simulation tools [21]. They offer preliminary analysis of the new dataset together with a selection of well-known datasets (including ISCX 2012). The inference model is modeled against six key features, each weighted equally:

- Complete capture of audit logs of computer operating system, together with network packet traces
- Maximum number of possible attacks included
- Current attack behaviours
- Real-world normal traffic dynamics with operation timings and industry complexity
- Maintenance of cyber infrastructure performance during complete capture
- Ground truth information included to assist labeling process

This is a welcome initiative, however we should recognise that lack of uniformity and uncertain knowledge in cybersecurity mean that any such quantitative analysis will necessarily be imperfect. An observation we should also make here is that some of these attributes are subjective; for example we know that not all attacks are likely to be characterised at any point in time (i.e. there are known unknowns); it also remains uncertain as to whether all simulated data can reasonably be ranked of equal quality - simulation itself relies on the quality of it's inputs, which may be both variable and subjective.

Another important observation is that some of these variables are temporal, and dependent on deployment context. For example, we know that attack behaviours and traffic dynamics change over time, and relative distributions of normal and abnormal traffic - even the presence of certain attack types - will vary, depending on where in the infrastructure events are captured, the type of organisation hosting the infrastructure, and what proportion of traffic is itself encrypted or tunneled.

This raises another important consideration: it is not always clear to researchers how data was collected (or simulated) and how representative the deployment context is. For example, an Internet backbone, enterprise, and industrial network will exhibit significant differences in event distributions, attack types, and traffic dynamics. Each organisation will have a very different threat profile and risk appetite, and the efficacy of security controls can vary significantly. Given the relatively low proportions of anomalous events, and the small incremental gains being made in intrusion detection accuracy, it is important that researchers are given clear guidance on dataset context, ideally using a well understood taxonomy, and a clear set of 'meta-labels' identifying the deployment and traffic model types.

It would also be worth considering other important factors in such analysis, such as whether a dataset is adequately maintained and updated, whether meta-information such as flow summaries are available, whether it is possible to derived advanced feature inputs from the dataset that may be useful in identifying malware (such as flow asymmetry, burst characteristics, payload entropy etc.), and the degree of anonymisation performed (if any) on the data, and data sensitivities. These are all important considerations in understanding how useful a dataset is likely to be, and its quality

Where models produce rankings of datasets this can be invaluable to researchers, however we should consider that the threat, cybersecurity and infrastructure environments are undergoing continual change, and several 'gold standard' datasets have become obsolete relatively quickly, due primarily to lack of maintenance. The product of such models therefore need to be framed in the context of time, and the state of each dataset with respect to the cybersecurity landscape at the time of enquiry. Model results should be time-stamped so that rankings between datasets can be assessed in context.

2.4 Simulated verses Live data

There are three broad classes of methodology used to generate datasets suitable for intrusion research, and these are identified in Figure 2:

- Type-L: Live Dataset; the passive recording of live (i.e. real) network and log event data from real-world contexts, ideally a representative production environment.
- Type-H: Hybrid Dataset. Involves setting up an emulated environment (such as a laboratory, or virtual machine) with real or emulated systems, to approximate a subset of a real-world context, injecting threat events from either synthetic agents, commercial test tools, or systems running live malware.
- Type-S: Synthetic Dataset: typically involves a synthetic model of the environment and threat actors (for example using virtual machines or specialised modeling tools), or simulated generation of a dataset from a set of configuration files.

Obtaining good quality datasets from representative live environments is a major challenge in threat research, due to the prohibitive cost, complexity and resources required to create and maintain, and security and privacy concerns. It is therefore rare to find Type-L datasets in the public domain, and these datasets tend to become rapidly obsolete. These challenges can be even more acute for emerging applications such as IoT, sensor

networks, and smart cities - indeed the Royal Academy of Engineering cites data quality as one of the six major barriers to effectively optimise smart infrastructures, alongside data privacy [90].

It comes as no surprise therefore that the majority of public datasets include some degree of simulation: some are entirely synthetic [18, 22], others a hybrid of live and synthesised attacks [20, 92], with a few contain purely live (i.e. real) data [23]. Synthetic data may be created using a variety of methods, including the use of specially design event generators, commercial simulation tools, network scripts, or by running samples of malware within a specially contained environment - such as a virtual machine. Synthetic data typically represents a subset of the events one might reasonably expect to find in a live network trace; it is generally difficult and impractical to reproduce such a broad range of event sources with accuracy. Synthesised data may lack the subtle timing characteristics of live threat traffic, and may even miss important features of an attack, particularly where threat event generators are overly simplified or behavior is approximated. Live traffic may also exhibit so-called 'zero-day attacks' - malicious attack events that have not been fully characterised at the time of analysis, and these may have been be overlooked in a synthetic model design.

In the case of emerging developments such as smart cities and sensor networks, these are complex heterogeneous systems, including both conventional networked systems as well as large deployments of Internet of Things (IoT) devices and wireless sensor networks (WSNs), with include high volumes of dynamic heterogeneous data. Pure simulation of such environments may not be ideal, and several recent studies have made use of live traffic from laboratory networks, together with a component of attack simulation (i.e. Type-H) [91, 92, 93, 94].

Simulation (Type-S data) does offer significant flexibility to experiment with many variables that would otherwise be difficult or impossible to explore with live traffic, and works best where it can be validated against real event profiles. Such datasets can be much easier to maintain (and even improve), and large complex systems can often be modeled, as well as a broad range of current threats. There are initiatives to create tools to help synthesis event data; for example [24] describes a packet-based data set generator (ID2T). This tool uses live benign events mixed with synthesised threats - inserted using scripts or by modifying Packet Capture (PCAP) trace files. Scripts were also used to create synthesised threats in the Coburg Intrusion Detection Data Sets (CIDDS) described in Section 3, with the associated python scripts published [25].

2.4.1 Important Tools

In the literature there are a number of open source and commercial tools and libraries frequently used to generate, analyse and capture attack and benign events, including: Nmap, MetaSploit, Splunk, tcpdump, tcppriv, Snort, WireShark, Bro, Argus, and PCAP. This is an evolving field and a maintained list with descriptions and links is provided at [16]: Specific exploit code is often available within the open source community. In several cases researchers have created custom tools to generate, simulate, and analyse exploits. It is strongly recommended that any testing of such exploits be performed using a mix of simulation models and heavily insulated (so called 'air-gapped') virtualised infrastructure to contain any resulting damage from malware. There are conveniently packaged virtual machines available to test malware, and these include many of the tools listed above; for example Security Onion [17] includes Bro and several other useful tools for isolation and analysis.

2.5 Raw Events, Metadata and Flows

Where possible datasets should include low level packet trace events and log information; referred to as a category I dataset in [26]. This raw event data enables datasets to be re-examined at any point in the future and are invaluable for correlation or deeper analysis, for creating flow summaries, and for creating useful metadata (such as summary statistics, counters, temporal and payload analysis). Whilst invaluable, discrete packet events represent a challenging data source from which to gain insight, because of the sheer volume of data, the granularity, dimensionality, and the lack of contextual association between events. As we discuss later, threat events typically form only a very small percentage of such datasets and it may be hard to separate the noise from any patterns of interest.

Arguably the most useful high level abstraction we can extract metadata from, in preparation for learning and statistical models, is that based on network event *flows*; referred to as category II datasets in [26]. Flows are essentially aggregated summaries of low-level logical connections (such as a user login session), and significantly reduce the data dimensionality to more manageable scale. For example, instead of analysing thousands of discrete HTTP events, it is far more useful to view these grouped by individual user sessions, with discrete start and stop times, source and destination IP address, and so on. Flows also bypass the problems of encrypted connections and therefore reduce privacy concerns. A widely used collection format is NetFlow, introduced by Cisco Systems [27], although there are several closely related industry formats and standards. The IP Flow Information Export (IPFIX) protocol, ratified by the IETF is 2013, is particularly attractive for intrusion research with machine learning, offering high flexibility in parameter mapping through extensible flow records [99].

Several of the datasets described here include both packet traces and flow summaries. [28] contributed one of the first flow-based data sets (described in Section 3). The authors collected flow-based data from a honeypot with several services and analysed the log files to label the corresponding flows. However, nearly all of the 14 million flows are malicious since real background traffic is missing. Wheelus et al. [29] and Zuech et al. [30] both proposed flow-based data sets (SANTA and IRSC respectively), though, neither of the two data sets are publicly available to the best of our knowledge. The UNB IDS datasets include useful flow summaries, as well as useful and detailed behavioural metadata summaries per flow. The CIDDS dataset makes extensive use of NetFlow summary data [26], [31]. Flow analysis has also been successfully employed together with machine learning in emerging contexts such as IoT [98].

A drawback of datasets comprising only high-level flow or metadata (i.e. where packet traces are absent) is that certain types of analysis cannot be performed. For example we might wish to analyse the payload of specific flows at a packet level in order to detect malware [32], or use payload to measure features such as entropy (e.g. to assess whether user data has been encrypted on a plaintext channel – suggesting a backdoor threat). A compromise would be to perform wider analysis on packet traces and expose additional metrics as part of the flow creation process. The UNB ISCX datasets for example expose considerable metadata in flow summaries, which reduces the need to examine raw trace files.

2.6 Detection Time and Accuracy

The ability to create well-tuned and accurate classification and predictive models is heavily dependent on the quality and volume of data available; timeliness and accuracy are critical features of intrusion detection and prediction systems. As we described earlier, if an attacker has privileged access to inside resources, the consequences can be very serious for the victim organisations; in terms of reputational damage, loss of sensitive intellectual property, disclosure of compromising or highly sensitive data, and direct and indirect financial loss. When we consider that the median time for detecting breach is currently in the order of 3-4 months [33] the inference is that many attackers have ample ‘dwell’ time to execute serious attacks and cover their tracks. We also know that much of the damage done in such attacks happens in a relatively short period after compromise. The inability to detect an intrusion quickly and accurately may also lead to regulatory failure. As an example consider the General Data Protection Act [4], which came into force in May 2018. This is a legal governance regulation that has the power to enforce heavy fines against organisations that do not a) provide adequate controls on Personally Identifiable Information (PII) data on EU residents, and b) notify in the event of breach in good time.

Intrusion and insider threat detection is a hard problem, and we have witnessed several generations of intrusion detection, prevention and correlation techniques deployed in commercial products over the past two decades, with mixed success. Attacks typically represent a very small percentage of total events, they can be executed using multiple techniques (so-called blended attacks), and the techniques are becoming increasingly sophisticated and stealthy. Given these points, IDS researchers have a number of particular challenges to consider when dealing with intrusion datasets, and dataset authors should ensure that these are fully covered:

- **Timespan**: Is the data representative for all significant load conditions and trends (e.g. peak loads, services that only run on specific days). Ideally we want to see data that spends *at least* a week of activity.
- **Timestamps**: Are all timestamps a) consistently recorded, b) synchronised across the capture domain (e.g. time drift accounted for, differences normalised across timezones), c) sufficiently granular, d) contiguous and consistent with actual event time (e.g. are some events batched)
- **Summary Records**: If the raw data has been summarised, are the nuances in flow timing adequately captured (e.g. have session time and burst characteristics been preserved).

Stealth attacks are a particular challenge, in the data may need to be recorded over very long time periods (potentially months), and this may be resource and cost prohibitive. We discuss this particular topic further in Section 4.1.3, where we offer some recommendations.

In real-world IDS deployments, one of the key issues that define success is the level of accuracy and timeliness with which a technique discovers threats, without raising false alarms (false positives) or classifying real threats as benign (false negatives). Inaccuracies in these techniques can quickly lead to operator fatigue (where operations staff become overloaded with false alarms), loss of customers (for example where a bank repeatedly issues fraud alerts to customers, for legitimate transactions), or a false sense of security (where false positives are suppressed to the point where fraudulent transactions are taking place undetected – i.e. a false negative). It is important therefore that we improve techniques to detect such intrusions quickly, whilst maintaining high accuracy. This requires the invention of novel techniques and wider access to large representative datasets. Given the relative inequalities in attack and benign traffic (with attack volumes of less than 0.0001% of overall events for example) researchers also need labeled data (for training) and rich feature sets with which to work.

The rarity of specific attack classes within such large event populations presents a major challenge to researchers. In the author's own analysis of the UNB-ISCX-IDS 2012 dataset (where events were recorded for an entire week) several attack vectors represent less than 0.00006% of total traffic. This can be misleading: importantly such attacks are often bursty – highly localised with respect to time (i.e. malware packets may be tightly clustered within short bursts, and therefore any averaging across a wider timeframe can be misleading in terms of relative percentages). Anomaly detection techniques may therefore benefit from including a temporal dimension, and by filtering on specific flow types (for example, by just examining web traffic). Threat event frequencies are likely to be much more significant within the localised time ranges associated with a specific flow type, In the UNB-ISCX-IDS 2012 dataset for example the average frequency of HTTP web threat events is 9.6%, however if we examine data at a more granular level the frequency within one of the event traces captured from day four is as high as 51.8%. If the event flows were filtered to isolate just web traffic then the relative proportions of threat events to normal events would certainly be higher.

2.7 De-Identification

Arguably the largest barrier to publishing representative datasets concerns privacy [20], with many organisations concerned that data (such as packet traces and logs) might reveal sensitive information on users, as well as infrastructure details; thereby giving malicious actors inside information which can later be exploited. There are risk and liability concerns, particularly with PII, in a climate of increasing punitive regulatory controls.

Superficially we might consider removing, masking or substituting names, addresses and identification numbers in order to improve data privacy, and in fact this is often done using a process referred to as anonymisation [20] or *de-identification* [34]. The problem with de-identification is that it either hides too much information, severely limiting any analysis, or gives away too much - where multiple correlated queries can reveal important personal or infrastructure details - perhaps even the organisation's own identity, in a process is referred to as re-identification.

It can also be technically difficult to perform de-identification where large raw packet traces are included, and where the organization needs some certainty on the lack of data leakage (for example payload data within packets may also include sensitive information missed by naive anonymisation procedures). De-identification

has been found inadequate in the face of modern data science techniques and computing resources [35, 36, 37]. The Fundamental Law of Information Recovery informally states that "overly accurate" estimates of "too many" statistics completely destroys privacy [38, 39]. There are techniques to attempt to mitigate this risk. For example *differential privacy* is a mathematically rigorous definition of privacy focussed on privacy leakage revealed by the analysis of large datasets; it includes a formal measure of privacy loss [39, 40, 41]. In effect differential privacy provides a mathematical guarantee that the data contributor will not be affected, adversely or otherwise, by allowing their data to be analysed, and includes a threshold specification for desired privacy loss as part of the input. At the same time sufficient information should ideally be retained to make analysis still relevant and useful.

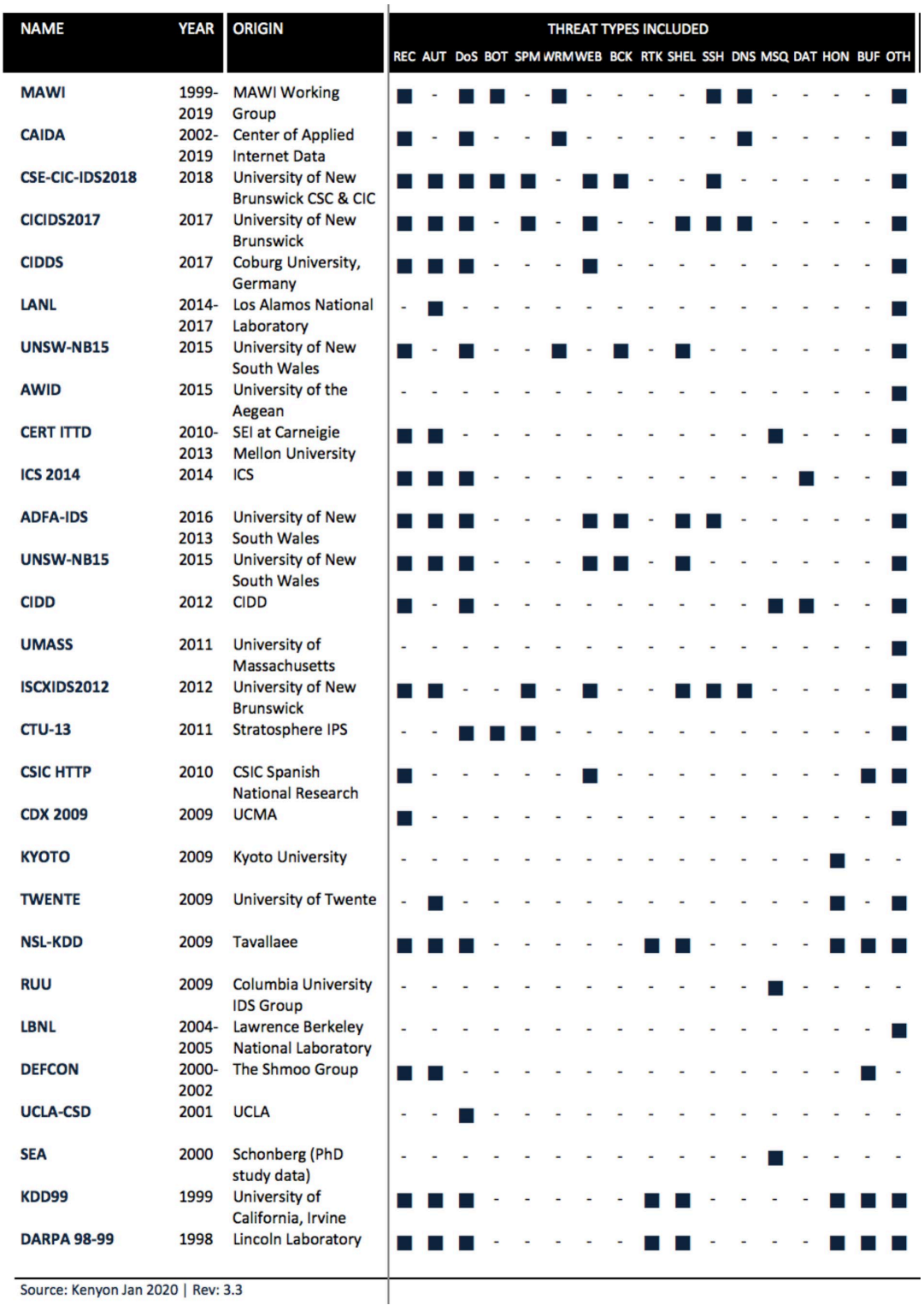
The reality is that anonymisation is hard to achieve on such datasets without the risk of data leakage. Modern data science techniques make it virtually impossible to achieve acceptable levels of privacy without severely weakening any subsequent analysis. Some of the datasets we discuss later in this report include heavy de-identification, and there are criticisms (discussed in Section 3) that this compromises - or even invalidates - analysis: limiting the utility of such a dataset. It is not unusual for example to see complete removal of all packet payload data from packet traces - since this removes any doubt of exposing encapsulated user data. The exclusion of payload data may limit useful analysis (such as calculating entropy values for the data in packet flows). The use of flow summaries, including useful metadata that is added prior to de-identification, is an excellent way to reduce privacy concerns, reduce data complexity, and support better analytics.

2.8 Legal and Ethical Concerns

A particularly challenging area of intrusion analysis is the collection and publication of datasets representing insider threat. There are several major obstacle, in particular 1) whether organisations are able to run or even simulate live insider threat, 2) the ability to fully define the scope of all threats, and perhaps even more challenging 3) the legal and ethical concerns raised when monitoring live user traffic. It is notable that there are very few datasets available that are specifically designed for such analysis, and these have generally been created in highly controlled environments, which are not representative of real-world insider threat [42, 43].

3 Publicly Available Intrusion Datasets

Earlier we discussed some of the key challenges in designing and maintaining public intrusion datasets; highlighting the problems in maintaining real-world representative data, and the speed of change, sophistication and volume of cybersecurity threats. As a consequence many researchers may be unwittingly working with sub-optimal datasets, with research often undertaken 'silos' across government, commercial, and academic institutions. In this section we attempt to characterise the most widely known intrusion datasets, providing comparisons of threat profiles, method of creation, and highlight the presence of important metadata and summary analytics. We discuss the general applicability of each dataset to research today, and note how well maintained these datasets are. Datasets are ordered by year of publication - or last update - whichever is more recent. As a general rule more recent datasets tend to be more useful to security researchers. Older datasets can still be valuable, for example when comparing new detection techniques with legacy models, however any dataset older than 2-3 years may be missing important threat vectors and distribution trends.



NAME	YEAR	ORIGIN	THREAT TYPES INCLUDED																
			REC	AUT	DoS	BOT	SPM	WRM	WEB	BCK	RTK	SHEL	SSH	DNS	MSQ	DAT	HON	BUF	OTH
MAWI	1999-2019	MAWI Working Group	■	-	■	■	-	■	-	-	-	-	■	■	-	-	-	-	■
CAIDA	2002-2019	Center of Applied Internet Data	■	-	■	-	-	■	-	-	-	-	-	■	-	-	-	-	■
CSE-CIC-IDS2018	2018	University of New Brunswick CSC & CIC	■	■	■	■	■	-	■	■	-	-	■	-	-	-	-	-	■
CICIDS2017	2017	University of New Brunswick	■	■	■	-	■	-	■	-	-	■	■	■	-	-	-	-	■
CIDDS	2017	Coburg University, Germany	■	■	■	-	-	-	■	-	-	-	-	-	-	-	-	-	■
LANL	2014-2017	Los Alamos National Laboratory	-	■	-	-	-	-	-	-	-	-	-	-	-	-	-	-	■
UNSW-NB15	2015	University of New South Wales	■	-	■	-	-	■	-	■	-	■	-	-	-	-	-	-	■
AWID	2015	University of the Aegean	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	■
CERT ITTD	2010-2013	SEI at Carneigie Mellon University	■	■	-	-	-	-	-	-	-	-	-	-	■	-	-	-	■
ICS 2014	2014	ICS	■	■	■	-	-	-	-	-	-	-	-	-	-	■	-	-	■
ADFA-IDS	2016 2013	University of New South Wales	■	■	■	-	-	-	■	■	-	■	■	-	-	-	-	-	■
UNSW-NB15	2015	University of New South Wales	■	■	■	-	-	-	■	■	-	■	-	-	-	-	-	-	■
CIDD	2012	CIDD	■	-	■	-	-	-	-	-	-	-	-	-	■	■	-	-	■
UMASS	2011	University of Massachusetts	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	■
ISCXIDS2012	2012	University of New Brunswick	■	■	-	-	■	-	■	-	-	■	■	■	-	-	-	-	■
CTU-13	2011	Stratosphere IPS	-	-	■	■	■	-	-	-	-	-	-	-	-	-	-	-	■
CSIC HTTP	2010	CSIC Spanish National Research	■	-	-	-	-	-	■	-	-	-	-	-	-	-	-	■	■
CDX 2009	2009	UCMA	■	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	■
KYOTO	2009	Kyoto University	-	-	-	-	-	-	-	-	-	-	-	-	-	-	■	-	-
TWENTE	2009	University of Twente	-	■	-	-	-	-	-	-	-	-	-	-	-	-	■	-	■
NSL-KDD	2009	Tavallaee	■	■	■	-	-	-	-	-	■	■	-	-	-	-	■	■	■
RUU	2009	Columbia University IDS Group	-	-	-	-	-	-	-	-	-	-	-	-	■	-	-	-	-
LBNL	2004-2005	Lawrence Berkeley National Laboratory	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	■
DEFCON	2000-2002	The Shmoo Group	■	■	-	-	-	-	-	-	-	-	-	-	-	-	-	■	-
UCLA-CSD	2001	UCLA	-	-	■	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SEA	2000	Schonberg (PhD study data)	-	-	-	-	-	-	-	-	-	-	-	-	■	-	-	-	-
KDD99	1999	University of California, Irvine	■	■	■	-	-	-	-	-	■	■	-	-	-	-	■	■	■
DARPA 98-99	1998	Lincoln Laboratory	■	■	■	-	-	-	-	-	■	■	-	-	-	-	■	■	■

Source: Kenyon Jan 2020 | Rev: 3.3

Figure 3: Attack features of public event datasets. Threat vectors include: reconnaissance (REC), authentication (AUT), Denial of Service (DoS), spam (SPM), worms (WRM), cross site scripting (XSS), SQL injection (SQL), backdoor (BCK), root kits (RTK), shell (SHEL), secure shell (SSH), masquerade (MSQ), data manipulation (DAT), honeypots (HON), domain (DNS), buffer overflow (BUF) and other (OTH).

3.1 MAWI Dataset (2000-2019)

The Measurement and Analysis on the WIDE Internet (MAWI) Working Group is collaboration between Japanese network research and academic institutions with corporate sponsorship, and focuses on traffic measurement and analysis; in particular the long-term measurement on wide-area, global Internet. This group has been working on building a large public network traffic dataset for use in research, collected over many years. The MAWI repository provides up-to-date packet-based data sets, created by capturing the network events from a number of sampling points for Internet wide area backbone traffic. Events are labelled by combining various anomaly detectors. The labels are obtained using an advanced graph-based methodology that compares and combines different and independent anomaly detectors [23]. The data includes useful summaries showing protocol distributions; together with corresponding of PCAP traces, captured using *tcpdump* (each around 7 GB in size). The archive contains labeled traffic anomalies. IP addresses in the traces are de-identified by a modified version of *tcpdpriv*. Some filtering is performed (ICMP) to remove noise. The data set is daily updated to include new traffic from upcoming applications and anomalies. PCAP files are available on a daily basis, as well as labeled XML and CSV summaries. The archive employs two distinct anomaly classification techniques based on protocol header features; protocols state information, and connection patterns A taxonomy of anomalies in wide are backbone traffic is proposed in [14]. Threats included in the dataset cover a broad range of backbone events such as DoS, port scanning, and other anomalies: for further information see Figure 3.

The MAWI datasets are well maintained, and contain live and labeled event data, however their broader utility is limited for training and evaluation of intrusion detection techniques for domains other than backbone networks, since the characteristics of Internet backbone traffic - both threat and benign – differ significantly. Nevertheless this represents an excellent resource for researchers wishing to examine long-term trends and large scale attacks such as denial of service on backbone networks.

3.2 CAIDA (Center of Applied Internet Data Analysis, 2002-2019)

CAIDA consists of several Internet backbone datasets [44]. Most of CAIDAs datasets are specific to particular events or attack types associated with wide area IPv4 and IPv6 backbone links (primarily volumetric DDoS and DNS attacks, port scanning, worms etc.). Payload, protocol, and destinations are heavily anonymised. This dataset is interesting (for large scale DDoS threat analysis for example), however event distributions are fairly specific to the network context (similar to the MAWI dataset), and a number of shortcomings have been outlined in [19, 20, 45, 46, 47, 48].

3.3 UNB IDS (University of New Brunswick, 2012, 2017, 2018)

The University of New Brunswick (UNB) has provided a range of invaluable well-labeled and well-maintained datasets for cybersecurity research [49,88], including several discrete datasets: ISCX IDS data set, ISCX Botnet data set, ISCX Android validation data set, ISCX Android Botnet Dataset. ISCX IDS offers realistic traffic (both attack and benign) that is reasonably current, and can be easily used to validate simulation models with labeled flow summaries. The IDS datasets are substantial, and incorporate both raw packet events as well as labeled flow summaries, captured over multiple days.

- **IDS 2012**: The 2012 IDS dataset (referred to as 'ISCXIDS2012') includes XML flow summaries with very useful metadata. The dataset has a number of limitations: specifically, the distribution of the simulated attacks is not based on real world statistics [20]. The dataset also contains a large percentage of web traffic, but does not include secure web (HTTPS) flows at the level that we would expect today.
- **IDS 2017**: The 2017 dataset (referred to as 'CICIDS2017') improves significantly on the 2012 release, and includes CSV flow summaries with a large number of additional properties exposed in the flow summaries. Real traces were analysed by the authors in order to create profiles for synthetically generating HTTP, SMTP, SSH, IMAP, POP3 (email), and FTP traffic, for 25 users. The range of threat events includes: Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and DDoS. The dataset comprises network traces with full packet payloads (in PCAP format).

- **IDS 2018**: The most recent (2018) IDS dataset (referred to as 'CSE-CIC-IDS2018 dataset for AWS') is a collaborative project between UNB and the Canadian Institute for Cybersecurity (CIC). The target organisation comprises 420 machines and 30 servers, logically organized into 5 departments, with the following protocols included: HTTPS, HTTP, SMTP, POP3, IMAP, SSH, and FTP (the majority of traffic comprising HTTP and HTTPS). The attack infrastructure comprises 50 machines, and seven threat scenarios are covered, including: Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and network infiltration (via a malicious email attachment). Note that the DDoS events included more modern application attacks (e.g. Slowloris). The dataset includes network traces and system logs for each machine, together with 80 features extracted from the captured traffic. The download size for this dataset is approximately 220GB. It requires an AWS client to download [88].

The UNB datasets are an excellent well-maintained resource for cybersecurity research today. Traffic generation is achieved using a specially designed flow generator called CICFlowMeter. Behavior is abstracted through configuration profiles (B-Profiles and M-Profiles for user and threat events respectively). Considerable thought has gone into the design and the statistical distribution of events, to be a realistic as possible. Note that some of the raw packet files are extremely large. Detailed analysis has been performed by the author on this dataset and a summary of the findings presented in Section 5.

3.4 CIDDS (2017)

CIDDS (Coburg Intrusion Detection Data Sets) is a labeled network flow-based dataset created in a virtualised OpenStack environment and NetFlow captures, CIDDS is designed for the evaluation of anomaly-based network intrusion detection systems. The main objective of CIDDS is the generation of customisable and up-to-date data sets, using unidirectional flows. There are two datasets: CIDDS-001 and CIDDS-002. The CIDDS dataset makes extensive use of NetFlow summaries [26], [31]. CIDDS-001 is designed to simulate a small business environment. Python scripts were used to simulate benign user behaviour (e.g. web browsing, email, and file transfer) within typical user schedules [25]. The dataset was captured over four weeks and contains nearly 32 millions flows: 31 million captured within OpenStack, and 0.7 million captured at an external server. The dataset includes 92 attack types (70 executed within the OpenStack and 22 targeted the external server). CIDDS-002 is a similar to CIDDS-001, but includes management data sources on a separate internal network, and additional external servers: the dataset includes 43 attacks collected over two weeks. To generate threat events, DoS, Brute Force and Port Scans were executed within the network. Since origins, targets, and timestamps of the executed attacks are known, labeled NetFlow data is also included. The external server was publically accessible for file and web traffic and was exposed to real threat events from the Internet. Only the publically accessible external IP addresses were anonymised. This is a potentially useful dataset for research, albeit with a relatively narrow deployment context and limited threat scenarious.

3.5 LANL (Los Alamos National Laboratory, 2014, 2015, 2017)

The Los Alamos National Laboratory (LANL) has released three data sets for public use [50, 51, 52]. The most recent, the Unified Host and Network Data Set is a subset of network flow and computer event logs collected from the LANL enterprise network over the course of approximately 90 days, in CSV format. The previous two datasets are essentially earlier versions of this dataset. The flow records included in the 2014 and 2017 archives are largely derived from internal routers, and the event logs from MS-Windows machines. The datasets have been de-identified: although in some cases important values were not de-identified (such as well-known network ports, system-level usernames and core enterprise hosts). [52] includes graphs to illustrate the quality of the network flow data set over time. The log events are useful to correlate where network traffic may be encrypted, and includes authentication events. The flow records are useful, but there are no PCAP traces, and the traffic distributions may not be ideal for developing intrusion techniques for other domains.

3.6 ADFA IDS (University of New South Wales, 2016)

The ADFA dataset [97] comprises several datasets created by the Australian Defence Force Academy (ADFA) in University of New South Wales (UNSW). The nomenclature, descriptions and provenance of these datasets are somewhat ambiguous in the literature, since ADFA includes several discrete datasets, with variants for Linux and Windows formats:

- NGIDS_DS
- ADFA: itself three datasets (ADFA_LD, ADFA-WD, and ADFA-WDSAA)
- Netflow_ids_label dataset.

The ADFA dataset emulates a host based intrusion detection (HIDS) systems and comprises of normal and abnormal Linux based system calls traces (with attack, training and validation log data). Since this dataset is limited to host logs and employs an artificially simple test-bed, this data of limited research value today; [19] describes several shortcomings. The Netwflow IDS dataset contained labeled Netflow events and is primarily aimed at network IDS (NIDS) evaluation. The event data is a collection of flow summary logs, collected over a five week period during 2012, and also has limited value.

NGIDS-DS is a more substantial dataset (approximately 6.73GB uncompressed), comprising extensive labeled network and host events (i.e. host logs and PCAP files, dated 2016), together with 'ground truth' and feature descriptions. The dataset is well described in [19], as part of a study to produce a uniform metric for quantifying intrusion dataset quality, and an effort to produce more realistic IDS data. Similar to UNSW-NB15, the dataset is synthetically produced using a commercial simulation tool (IXIA PerfectStorm), used to generate a mixture of normal and abnormal traffic. The abnormal traffic comprises seven major attack classes, including Exploits; DoS; Worms; Generic; Reconnaissance; Shellcode; and Backdoors. In [19] the authors state that the NGIDS-DS dataset achieves "medium-high quality" realism, measure using a qualitative modeling approach based on a Sugeno fuzzy inference model, primarily due to the setup and use of the IXIA security test hardware.

3.7 UNSW-NB15 (University of New South Wales, 2015)

The dataset is created by by the Australian Defence Force Academy (ADFA) in University of New South Wales, The dataset is synthetically produced using a commercial simulation tool (IXIA PerfectStorm) in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) [95, 96]. The dataset contains relatively recent distributions of normal (live) and attack traffic (synthesised). The dataset represents activities found in critical infrastructure across enterprise networks. It includes 9 attack types (Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worm). Tcpdump was used to capture 100 GB of the raw traffic, with the dataset comprising 257,705 records labeled as an attack type or normal. The training set comprises 175,341 records. The test set holds 82,332 records, 16,353 of which correspond to DDoS attacks. Both Argus and Bro-IDS flow and analysis tools were used, with 12 algorithms implemented to generate 49 features, with a class label split into five feature sets: flow, basic, content, time, and additional generated. Note that the DoS attack traffic is largely based on Reflectors (a legitimate computer controlled by an attacker); hence source IP addresses cannot be used to classify DoS traffic. The data is well structured, well described, and well labeled [96], and is more complex than many previous datasets, making it a useful benchmark to evaluate, however [19] describes several shortcomings in both the design and threat methodology.

3.8 Aegean WiFi Intrusion Dataset (AWID), 2015

The Aegean WiFi Intrusion Dataset (AWID), a publicly available dataset containing a rich blend of normal and attack traffic against 802.11 networks [91]. AWID contain traces of live 802.11 traffic for both normal and threat events. The event traces were extracted from a dedicated Wireless Encryption Protocol (WEP)

protected 802.11 networks; although the environment comprised a laboratory emulation of a Small Office Home Office (SOHO) Wi-Fi network, and a single attacker was used to generate threat events. Note that the threat events restricted to WEP attacks and vulnerabilities. AWID comprises two labelled sets (AWID-CLS, AWID-ATK), each collected over one hour, with the second follows a more detailed labelling classification based on the actual 16 attack types. An extensive evaluation of AWID using several machine learning algorithms is provided in [92]. AWID is unusual in that it is focused primarily on use in intrusion detection in Wi-Fi networks, and it believed to be the first publicly available dataset of this kind [92]. As such AWID represents one of few public datasets (particularly ones that contain live traffic) in this field, and therefore remains a useful source for research, despite its age and the limitation that is contains only synthesised WEP related attacks from a single source.

3.9 Industrial Control System (ICS) Cyber Attack Datasets (Mississippi State University, 2014)

This dataset is fairly unique in offering an IDS dataset for critical infrastructure research in industrial control systems [89]. It features events created with the Supervisory Control and Data Acquisition (SCADA) protocol, generated by the Mississippi State University's critical Infrastructure protection center, in cooperation with Oak Ridge National Laboratories (ORNL). The dataset is created synthetically and randomly sampled. There are five distinct datasets:

- **Dataset 1**: power systems data, with measurements related to electric transmission system normal, disturbance, control, and cyber attack behaviors. Includes synchrophasor measurements and data logs from Snort, a simulated control panel, and relays
- **Dataset 2**: Gas pipeline Modbus data. ORNL format.
- **Dataset 3**: a database of cyber attacks against 2 laboratory scale industrial control systems; a gas pipeline and water storage tank. WEKA format.
- **Dataset 4**: New gas pipeline data, created to mitigate design flaws in the earlier dataset (primarily to improve randomness). ARFF format.
- **Dataset 5**: Energy Management System (EMS) log data, anonymised and captured over 30 days

The datasets include a set of 28 distinct attacks on these systems, comprising many attacks that would not be encountered in other domains (e.g. attacks on relays and the Modbus protocol). The datasets have been analysed for power system cyber attack classification and a list of associated papers can be found at [89]. Whilst this dataset is now effectively quite old, its novelty and the range of specialised events are certainly worthy of examination for IDS researchers.

3.10 CERT ITTD (Carnegie Mellon University, 2010-2013)

The Software Engineering Institute (SEI), part of The CERT Coordination Center, created by the Defense Advanced Research Projects Agency (DARPA) in 1988, produces a collection of synthetic insider threat test datasets called the Insider Threat Test Dataset (ITTD). These datasets provide both synthetic background data and data from synthetic adversaries [18]. Datasets are structured based on the release of the data generator that created them; the data generator is able to simulate human behavior on networked systems, and is described in [22]. At the time of publication release 1 through 6.2 were available, with later release normally including a superset of the functionality in earlier versions. The release dates are unclear, but from the data files these appear to have been generated between 2010 and 2013. The data comprises mainly log event data (including login sessions, LDAP identities, web, file and email records). There are no PCAP or flow records. This dataset is interesting, in that it attempts to simulate realistic user behavior (including insider threats), but nevertheless quite specialised. There are some deficiencies in the datasets (for example user login temporal distributions and threat event distributions may not be entirely realistic). Evaluation of the synthesis techniques also suggested a need to explore more sophisticated models [18].

3.11 ADFA-IDS (University of New South Wales, 2013)

This dataset includes normal, training, and test data [21]. To create the ADFA dataset, authors installed an Apache web service, together with a MySQL database and FTP servers, Tikiwiki content management system, and remote access. The dataset includes a limited set of attack vectors including FTP and SSH password bruteforce, Java based Meterpreter, Linux Meterpreter payload and C100 Webshel. The dataset has a number of limitations,: there is a lack of variety in threat types, and some of the attack behavior is not well separated from normal behavior [53, 54]. The value of this dataset to researchers is therefore limited.

3.12 Cloud Intrusion Detection Dataset (CIDD, 2012)

CIDD is the intrusion dataset for masquerade analysis in cloud systems [55], and comprises both knowledge and behavior based audit data, collected from a mix of UNIX and Windows users (128 in total). CIDD includes real instances of host and network based attacks and masquerades, and provides complete diverse audit parameters to build efficient detection techniques. Four earlier datasets have traditionally been used to evaluate masquerade detection techniques: SEA [43], the Greenberg dataset [56], Purdue (PU) [57], and RUU [42]. These datasets suffer from several limitations. CIDD attempts to resolve these deficiencies with respect to cloud contexts, and covers a much larger user population, with additional audit data. CIDD includes features to detect masquerade attacks, and more than 100 additional attack types, including: Denial Of Service (DOS), User to Root (U2R), remote to user, reconnaissance, data manipulation, and anomalous user behavior. CIDD includes both knowledge and behavior-based audit data. To build CIDD, the authors implemented a Log Analyzer and Correlator System (LACS) [55], to extract and correlate user audits from a group of log files in both host and network environments. Given the age of this dataset, limited information on setup, and that it appears to be derived from earlier DARPA dataset (based on a government network), the value is questionable.

3.13 UMASS (University of Massachusetts, 2011)

The dataset includes trace files for both network and wireless infrastructures [58, 59], generated using a single TCP-based download request attack scenario. The dataset is not recommended for IDS or IPS research due to the lack of variety of traffic and attacks, and the lack of sophistication in threat design [60].

3.14 CTU-13 Dataset (2011)

The CTU-13 Dataset [87], is a labeled dataset captured at CTU University (Czech Republic), and contains benign events, as well as threat events from a variety of malware. It includes live Botnet attacks (Spam, DDoS, Clickfraud etc.), with flow records (Argus, NetFlow), logs and PCAP captures. Labelling of threat events is based on the IP addresses used by the botnets. [20] proposed methods to generate labelled flow-based data sets for IDS, and describe various threat and benign event profiles which can be combined into new data sets. The data sets published by [20] and [61] contain bidirectional flows. Since the dataset is now several years old the lack of more recent threat events will limit its use.

3.15 CSIC HTTP (2010)

The CSIC 2010 HTTP dataset [62] was created by the Information Security Institute of CSIC (Spanish Research National Research Council), and was developed in response to the criticisms of earlier public datasets, the lack of test sets to verify Web Application Firewall (WAF) behaviour, and the privacy challenges inherent when using live data (although it appears that some randomised real user data is included during synthesised event generation). The dataset is designed for testing web-based attacks and comprises automatically generated and labeled HTTP requests (36,000 normal and over 25,000 anomalous). It includes simulated attacks such as SQL injection, buffer overflow, CRLF injection, Cross Site Scripting (XSS), server side include, as well as attempts at information gathering, file disclosure, and parameter tampering. The dataset included normal (test and training) events as well as anomalous test events, and has been used with some success for web based threat detection [63, 64, 65, 66, 67, 68]. The events are provided as limited CSV

summary metadata, and lack of HTTPS traffic makes this dataset unrepresentative of current web traffic, and therefore of limited value in IDS research.

3.16 CDX 2009 Log Dataset (United States Military Academy, 2009)

The CDX 2009 log dataset contains approximately 19 MB of log and alert data captured from UCMA CDX network, connected to untrusted Windows machines provided by the National Security Agency (NSA). This includes: a) Snort IDS Alert log (four days of alerts). b) DNS Logs: four days of events, including external 'named' events , and DNS messages. c) Apache Webserver Logs: twenty-four hours of events, including an access log, and error log. d) Splunk Log Server Aggregate Log: nine hours of events. The data is essentially unlabeled, but includes timestamps, and the Snort IDS Alert log provides some alert information that could be correlated in lieu of labeled data. In this dataset common attack tools namely Nikto, Nessus, and WebScarab have been used by attackers to carry out reconnaissance and attacks automatically. Benign traffic includes web, email, DNS lookups, and other required services. The CDX dataset demonstrates how network warfare competitions may be useful in generating labeled datasets. The dataset could be used to test IDS detection techniques, however the environment is artificially constructed (by design) to include a high level of threats, and so it suffers from appropriate traffic diversity to be directly transferrable to other domains [69].

3.17 Kyoto (Kyoto University, 2009)

This dataset has been created using honeypots, so there is no process for manual labeling and anonymisation, furthermore it has limited view of the network traffic because only attacks directed at the honeypots can be observed. It has ten extra features such as IDS_Detection, Malware_Detection, and Ashula_Detection than the previous available datasets, which are useful in NIDS analysis and evaluation. Since normal traffic is simulated repeatedly during the attacks and only produces DNS and mail traffic data - which does not reflect real world normal traffic - there are no false positives. The absence of false positives is important as it minimises the number of alerts [70, 71, 72]; as such the value of this dataset for research today is limited.

3.18 Twente (University of Twente, 2009)

This is one of the earliest datasets to include flow based summaries; although almost all of the 14.2 million flows collected are malicious since real background traffic is missing. Approximately 98% of the flows are labeled [28]. To create this dataset, three services OpenSSH, Apache web server and Proftp using auth/ident on port 113 were installed to collect data from a honeypot network using NetFlow. Some side-effect traffic (such as auth/ident, ICMP, and IRC traffic - which are not completely benign or malicious) are included. Moreover, it contains some unknown and uncorrelated alerts traffic. The data is more realistic than many prior datasets, however the lack of volume and diversity of attacks are major limitations. [73], and this dataset is not recommended for IDS research purposes.

3.19 NSL-KDD (2009)

In order to resolve some of the key shortcomings in the KDD CUP'99 (and underlying DARPA'98) datasets [74] proposed the NSL-KDD dataset [75]. NSL-KDD attempts to mitigate several weaknesses in the original dataset by further processing the original event: for example by reduction of attack event duplicates by 88%. Despite these improvements NSL-KDD still suffers the same problems highlighted by [76], and the value of all DARPA-derived datasets is questionable: given that this data is nearly two decades old and seems unlikely to contain event distributions (threat or benign) that are representative of today's deployment contexts. For current IDS research purposes this dataset is now outdated and should be considered obsolete.

3.20 RUU IDS Dataset (~2009)

The RUU (pronounced Are You You?) dataset is derived from the Columbia IDS Insider Threat Detection project [42]. The aim of this project was to create technologies aimed at monitoring and detecting malicious insider activity in the context of host based systems; overcoming some of the limitations of the SEA

masquerade dataset [43]. The dataset tracks the activities of 34 users, and was created under tightly controlled and artificial conditions, to normalise testing, reduce variables and remove bias. The role of the malicious actor for example is orchestrated by essentially setting scripted objectives for some users, who were then provided access to unlocked computers of colleagues for a limited time. This is not how adversaries generally operate from several perspectives. RUU provides useful insight given the scarcity of representative masquerade data, however it's value today is limited given its age, and contrived methodology used to simulate maliscious activity: it should therefore be considered obsolete.

3.21 LBNL (Lawrence Berkeley National Laboratory and ICSI, 2004-2005)

LBNL's dataset of packet traces are full header network traffic recorded at a medium-sized internal enterprise network. The dataset excludes payload information and suffers from a heavy anonymisation to remove any information that could identify an individual IP [77]. At the time of publication this dataset appears to be no longer available. For current IDS research purposes this dataset is now outdated and should be considered obsolete.

3.22 UCLA CSD Packet Trace Dataset (2001)

The UCLA CSD packet trace database [78] contains TCP, UDP and other packet traces (including attacks) collected from a border router in the Computer Science Department, University of California Los Angeles. Simulated volumetric DDoS attacks are present. The events are part anonymised. Given the age and deployment context and unsophistication of the attack methods, for current IDS research purposes this dataset should be considered obsolete.

3.23 DEFCON (The Shmoo Group, 2000)

The DEFCON-8 dataset comprises port scanning and buffer overflow attacks, whereas the DEFCON-10 dataset (created in 2002) includes port scanning, bad packets, administrative privilege, and FTP by telnet protocol attacks [79]. This traffic was generated during the Capture the Flag (CTF) competition, and is therefore significantly different from the real world network traffic comprising mainly of attack traffic. This dataset is therefore useful in studying attacks and alert correlation techniques [79, 80]. If blended with simulated benign traffic at appropriate levels it could feasibly be used more widely, however the age and skewed event distribution in the dataset makes its utility today extremely limited. It should be considered obsolete for IDS research.

3.24 SEA Dataset (~2000)

This database is interesting and relatively unique, in that it provides command sequence data keyed in by real users logged into Unix systems, resulting in a sequenced dataset of 15,000 commands per user [43]. The first 5000 commands in each user command sequence represent normal actions. Thereafter, at various points within the command sequences, other users may be masquerading as the host user by inserting their own command sequences. It presents a particularly hard challenge to researchers in identifying 'foreign' session hijacking. A number of papers have been produced where statistical analysis has been used to detect variance in the key sequences and thereby identify the possibility that the user is not who they say they are. Whilst historically significant this dataset is of limited value today: temporal and geo-location information are absent (i.e. no session information), there are no command arguments, and the masqueraders are in effect simulated as session transpositions of other normal users.

3.25 ICS KDD '99 Dataset (University of California, Irvine, 1998, 99)

This data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, and is widely cited as a source of study for intrusion detection, and one of the most widely used datasets for NIDS evaluation. This dataset contains summarised flow data, together with a broad selection of intrusions simulated from a military network. 81 includes more than 20 attacks such as neptune-dos, pod-dos, smurf-

dos, buffer-overflow, rootkit, satan, teardrop, to name a few [81]. The KDD Cup 1999 dataset is now two decades old, created by processing the tcpdump logs from the DARPA'98 dataset. It therefore suffers from the same issues as the DARPA dataset, and has been heavily criticised by researchers as unrepresentative (on the distribution and mix of threats and anomalies, the high number of duplicates, and errors in the data) [74, 76]. The network traces of normal and attack traffic are merged in a simulated environment, resulting in a large number of redundant records that are littered with errors, leading to distorted test results [74]. Tavallaee et al. [74] highlights the large number of duplicate records, particularly with the threat events, finding 78% and 75% of the records are duplicated in the training and test set respectively. The large amount of duplicates in the training set will bias learning algorithms, inhibiting learning - especially on low frequency events (which paradoxically are often the more serious threats). Their work resulted in the introduction of a cleaned up variant of KDD'99 called NSL-KDD dataset in 2009 (described earlier). For current IDS research purposes this dataset should be considered obsolete.

3.26 DARPA IDS Dataset (Lincoln Laboratory, 1998, 1999)

Perhaps the most widely used dataset for intrusion analysis is the Intrusion Detection Evaluation Data Set provided by DARPA [82, 83, 84]. Currently three large-scale data sets are available: 1998, 1999 and 2000. These comprise offline and real-time tcpdump binary files and various audit logs, recorded on a daily basis over a number of weeks, with classifiers available for training. DARPA'98 comprises around 4 GB of compressed binary tcpdump network traffic, captured over 7 weeks. This represents about 5 million connection records, each averaging 100 bytes. This data has been widely used as the source data for the annual KDD Cup Challenge [81] (an open competition for security researchers to test techniques against a benchmark data set). Importantly, this data is itself generated under simulation conditions, and is a static repository. A number of sites participated in the DARPA off-line intrusion detection evaluation during 1999, where a test bed was used to create live background traffic similar to that on a large government site. The database holds 58 attack types (mixed with normal traffic), launched against UNIX and Windows NT hosts over several weeks. The dataset includes email traffic, web and FTP traffic, IRC, remote telnet, as well as SNMP monitoring of remote routers. Attack profiles include DoS, password guessing, buffer overflow, remote FTP, SYNflood, Nmap, and rootkit.

Although there are many papers citing research and describing this database, researchers have put forward multiple criticisms of DARPA, primarily due to issues associated with the artificial injection of attacks and benign traffic, lack of true representation of real-world network traffic, lack of attack data records, as well as irregularities (such as the absence of false positives) [76, 85]. McHugh's [76] critique is primarily based on the techniques used to generate the data set. Mahoney and Chan [86] found evidence of simulation artifacts in background traffic that could inflate the performance of some anomaly detection techniques. [74] highlights the high duplicate count, the potential for packet loss in the trace recording process, and that there are also no precise descriptions of the attacks themselves. For current IDS research purposes this dataset is now outdated and should be considered obsolete.

4 ANALYSIS

In this section we summarise our analysis on the datasets reviewed in Section 3; we highlight significant flaws and deficiencies in dataset designs, we describe features that should be considered best practice, and we identify specific datasets that demonstrate some of the features of best practice. We also clarify why IDS researchers should pay close attention to the provenance and composition of each dataset, and focus their attention on representative modern datasets in order to advance this field.

4.1.1 Flaws in Dataset Design

In reviewing these datasets - many of which are still in use today - we find significant flaws in their design, in the representativeness of threat and anomaly event types and frequencies, and in the clarity of labeling and description associated with many of the datasets reviewed in Section 3. Specifically:

- **Poor Provenance**: in many cases the provenance of each dataset is unclear (specifically: the date of event generation, the mode of event generation, and the environment in which the events were captured or created). It is important that researchers fully understand both provenance and context, and we have labeled each of these datasets with this information in Figures 2 and 3.
- **Lack of Maintenance**: the majority of dataset of not maintained, meaning that the threat events and distributions are unrepresentative of today's threat context, and have become obsolete. We have clearly identified where datasets have been well maintained and where they are obsolete or of limited use in Section 3. The scarcity of IDS datasets (especially for specific domains) may still tempt researchers into using flawed or obsolete data, and we caution against this.
- **Limited Origin Data**: many early datasets do not include origin data (such as raw packet captures, PCAP or TCPdump files). Often this means that important information is omitted from the event data (e.g. timestamps, flow direction, protocol flags etc.), and the lack of raw (original) event data means that these flaws cannot be rectified. This lack of critical features is a serious obstacle when comparing techniques across various datasets
- **Limited and Unverifiable Metadata**: many early datasets include only limited high level metadata, and important information is often omitted (e.g. timestamps, flow structures, event direction, protocol flags etc.). This is often compounded by the lack of raw origin data, with which it may be possible to verify and extend the original metadata. This lack of important features, and lack of ability to extend, verify or normalise those features is a major obstacle when comparing IDS techniques across datasets
- **Overly Simple Synthetic Models**: many early datasets are synthetic, created at a time when dataset design was relatively immature. These datasets often contain limited features, and significant design errors: such as duplicate events, and unrepresentative threat and anomaly event distributions. In some cases the original tooling may also be missing, and it is therefore difficult to recreate, verify or correct issues in such data.
- **Weak Dataset Availability for Insider and Masquerade Threats**: datasets representing insider threat and 'masquerade attacks' are particularly badly represented, and those datasets that are available are either obsolete or have multiple design flaws in their construction. This is primarily due to the privacy constraints around data capture, and the complexity in modeling the full scope of hacker behavior on a population of normal users. It is likely that these issues will only be resolved by more sophisticated synthetic models and/or major improvements on de-identification techniques.
- **Narrow Domain Contexts**: many of the datasets are synthetic, and are mostly representative of academic or government institutions. There are many important deployment contexts that are absent or poorly represented (e.g. financial networks, retail networks, industrial, modern cloud infrastructure). Traffic and threat characteristics vary significantly between these domains, and this lack of representation has the potential to hinder broader IDS research, since models may be insensitive to some of the features in such domains.
- **Sample Space too Short**: some of the datasets are effectively 'snapshots' of activity, and as such not adequately representative. Event distribution for both threat and normal events often exhibit strong temporal characteristics, and therefore datasets should include sufficient event captures, taken over multiple days, ideally at least a full week.
- **Limited Threat Coverage**: datasets rarely include a comprehensive set of threat events; this is partly due to the fact that datasets are frequently unmaintained, the constraints of the deployment context, and sometimes due to limitations in the infrastructure assets and resources available or overly-simplified synthetic tooling. We have labeled the distribution of threat events by dataset in Table X, enabling researchers to readily identify the datasets most appropriate.
- **Skewed Event Distributions**: anomaly and threat event frequencies are typically extremely low when compered to normal traffic on real networks. Several datasets contain much higher proportions of threat traffic (e.g. cyber-range related datasets) that would be expected. Further, modern encrypted web traffic is largely absent from many of the earlier datasets, which has major implications for IDS detection models (such encrypted traffic would obscure many features previously available).

Whilst many of these flaws become obvious with the benefit of hindsight, and we recognise that this field has matured significantly since the first datasets were published; nevertheless we must highlight these issues objectively, as some of these datasets are still be in use today due to the scarcity of IDS data and lack of

clarity around on their composition and design. We would also point out that many of these flaws may exist with other cybersecurity datasets to varying degrees.

4.1.2 Implications for IDS Researchers

As discussed earlier, the cybersecurity threat landscape has evolved substantially, and today we are seeing several important trends, including:

- A significant increase in the attack frequency and volume of certain attack types (e.g. volumetric DDoS and BotNet attacks).
- More sophisticated and stealthy attacks (long term reconnaissance, polymorphic malware, attacks increased compromised identity). Attacks that mimic real user behavior are becoming increasingly common and are typically much harder to detect.
- Much broader attack types, including geo-spatial features, discrete attack phases, diversionary attacks, mobile malware and machine and sensor based attacks.

In parallel, IT infrastructures have also changed markedly in recent years, including:

- The introduction of higher speed media types, new protocols and services
- A major shift towards encrypted traffic – especially web services
- Increased use of cloud and mobile services
- Machine and sensor generated data,
- Significant increases in network traffic, storage, and log event volumes.

These changes have had a dramatic effect on network usage patterns, and therefore more recent IDS datasets are likely to contain major differences in both the type and distribution of events present; in particular the distribution and volume of threat and anomalous events, as well as the range of services and protocols represented, Further, since modern networks typically carry a much higher proportion of encrypted data (such as HTTPS web traffic), pre-existing payload features may no longer be available to IDS detection models. The timing and geo-spatial characteristics of certain attack types may also differ significantly, as new threat types emerge (especially those mimicking real user behaviour).
In practice this means that there is questionable value in using older unmaintained datasets, especially where provenance is unclear, and the original raw event data is not available for scrutiny. When modeling IDS techniques it is critical that researchers use more recent datasets to ensure that important features are adequately represented, as these are likely to have a substantive effect on tuning and configuration (especially in machine learning applications). Since attack and anomalous traffic typically represents only a very small proportion of the traffic, models generated from unrepresentative data, are likely to prove ineffective in production environments. Researchers are strongly encouraged not to use those datasets labeled as ‘obsolete’ or ‘of limited value’ in chapter 3. In Section 4.1.4 we highlight some of the public datasets we do recommend for current research use.

4.1.3 Characteristics of Best Practice in IDS Dataset Design

In summarising the flaws in existing datasets we reach a set of recommendations on the key characteristics an IDS dataset should have, namely:

- Clear provenance on the dataset origin, publication date, deployment context, and any ethical points
- Broad and accurate representation of normal and threat events, appropriate to the deployment context
- Good quality timing and geo-spatial supporting data
- Archives of original data (such as PCAP traces) as well as high level summaries (e.g. flow records)
- Full disclosure of any methods of de-Identification used, with scope and implications.
- Clear documentation on any constraints and limitations in the design or data

- Publication of tools, source code and configuration data with which to build the dataset
- Supporting evidence of representational quality with respect to comparable live production environments (especially where event data is synthesised)
- A clear update path for future maintenance

Figure 4 provides a complete list of 14 features we recommend for ensuring future dataset quality, ranked in general order of importance. Note that this ranking is somewhat subjective and context-dependent, and is based on a number of factors, specifically we take the view that:

- More recent data is better than old data – in the context of cybersecurity this is an important concept.
- For some datasets, especially those created from live environments containing PII data, and highly regulated markets, aspects such as data provenance and consent may be extremely important. For synthetic data this may be less of an issue.
- We should be able to ascertain how a dataset has been designed, created, labeled and versioned, to avoid incorrect assumptions.
- We should be able to ascertain to which deployment context it represents (e.g. cloud enterprise, industrial), to avoid incorrect assumptions about the scope of threats.
- Accurate timestamps and appropriate time scope are almost always important for most IDS techniques, since this is largely a spatio-temporal problem. This may be less important for some techniques such as clustering.
- Flow summaries are highly useful to reduce dimensionality, however so long as the original raw event data is present in the archive we should be able to create these from source.
- If appropriate, any de-identification techniques used should be documented, to avoid inaccurate assumptions, and to understand what data may have been lost or obfuscated.

As we have discussed previously, ranking is not exact, and some compromises may have to be made by researchers if, for example, the dataset is the *only* representative for a particular domain (as we highlight in Section 4.1.4). If researchers are required to make such a compromise then they should consider the best practice criteria outlined here, and make reference to any omissions or shortfalls in their assumptions.

Designers of IDS datasets are strongly encouraged to take note of the features, and where possible integrate them in their design and publication processes. Since datasets may be constructed from a variety of techniques (including recording live network traffic, creating synthetic events, or hybrid of the two), and infrastructure scale and complexity may also vary widely, it may not always be possible to label all data or characterise the environment fully: where such limitations exist these should be clearly stated.

ID	PRI	FEATURE	DESCRIPTION
DP	M	DATASET PROVENANCE	Date of dataset creation and key authors. Clear description of the methodology used in the dataset design (i.e. simulated, live, or hybrid event generation). Links to archive repositories. A clear list of constraints or limitations in the dataset design and event scope.
DC	M	DOMAIN CONTEXT	Clear taxonomy and description of the deployment context represented in event distributions (e.g. cloud, industrial, academic, SCADA, cyberrange).
EC	M	ETHICAL CONTEXT	Where datasets iinclude sensitive user or organisational data, there should be a clear statement on whether appropriate consent (and/or other associated permissions) were sought, and provided. Evidence should ideally be published with the dataset. See also the point 12 on De-Identification.
CL	M	CONSISTENT LABELLING	The dataset should includes 'ground truth' (I,e, normal event) information, with consistent labeling of event types for both normal and threat events - where appropriate. If the dataset is recorded from a live production environment and does not include labels then this should be clearly stated in the dataset provenence and domain context.
EV	M	REPRESENTATIVE EVENTS	Includes up-to-date threat, anomaly and normal event distributions, with good coverage of recent protocol and service usage, and appropriate distributions of normal and threat events. Events patterns should be realistic and threat variety and volumes should be conssistent with the domain context.
SD	M	SAMPLE DURATION	Events should be sampled over a reasonable time period to demonstrate any associated variations in threat and normal event distribution patterns. For events recorded from live production environments at least one week of data should ideally be sampled, to ensure good coverage of the full range of services and demand spikes. Note that long range attacks may not be present even in such a large dataset.
TS	M	TEMPORAL SCOPE	Includes consistent timestamps with realistic levels of granularity. Includes temporal reaslism in event distributions that compare with real networks (e.g. business hours, seasonal changes). The specific timing of any included attacks (if known) should ideally be described.
SS	M	GEO-SPATIAL SCOPE	There should be support files documenting the main assets used in the dataset (servers, clients, threat sources, users etc.), including appropriate details on geo- and logical locations, including associated network addresses, domain names, physical and logical locations, and user and session identities (where appropriate)
MD	M	USEFUL METADATA	Well described metadata should ideally be present to assist in reducing dimensionality and in understanding high level abstractions. For example event summaries, statistical analysis, and flow summaries should be included. Where statistics and summaries are provided these should be in an industry standard format (e.g. CSV, XML, JSON etc.), and include details on methods used.
ES	D	CORRELATED EVENT SCOPE	Ideally there should be events recorded from LAN, WAN, and server enviornmnets (e.g. log files) that can be correlated both spatially and temporally against threat event sources, over time, and across various address spaces.
OD	D	ORIGIN DATA	Includusion of raw packet traces and audit log data that can be analysed and correlated independently of any metadata summaries and potentially processed at a later date
UP	D	UPDATE SUPPORT	The provision of tools, virtual machines, event generators, configuration templates and source code should where possible facilitate the updates and potential transformation of the dataset at a later date. This is significantly enhanced by including origin data and clear supporting documentation on the model infrastructure.
CD	D	CALIBRATION DETAILS	Details of calibration against realistic datasets in comparible live environments (or clear description of the reasoning behind skewed event distributions (e.g. cyberrange dataset)
DI	D	DE-IDENTIFICATION CONTEXT	Includes clear descriptions of any de-identification methods used (if applicable), including tools. Dataset should not be destructively anonymised to the point where context is lost.

Figure 4: Assessment criteria for intrusion dataset quality (where M=Mandatory, and D= Desirable). New datasets should attempt to include as many of these features as practical to accompany publication and assist researchers in assessing dataset suitability.

We also wish to point out that cybersecurity attacks may be particularly stealthy and take place over a long period of time (possibly many months). It may be infeasible or cost prohibitive to store or even process such large volumes of event data; however it could be critical for predictive or forensic analysis to have access to this data. We have already highlighted a difficulty in analysing raw event data, and one way to address this would be to reduce dimensionality and size by summarising these events using *flows* (introduced in Section 2.5). Flows are particularly attractive for use in machine learning models because of this dimensionality reduction. Where necessary, raw events could be stored offline in low cost storage. Alternatively, if bulk storage is not possible, it may be appropriate to store randomised samples of event data, representative of different time intervals, although this very much depends on the purpose of analysis and the level of audit mandated.

4.1.4 Recommended Datasets for IDS Research

Several of the datasets described in Section 3 do exhibit many of the features highlighted in Figure 4, and researchers are encouraged to consider these datasets for active research today. We specifically highlight a subset of the datasets below, the first two of which are substantially consistent with best practice guidance in Section 4.1.3. These datasets are maintained, the threat compositions are up to date, they have clear

provenance, are well-described and thoughtfully designed. These datasets also reflect the changes in composition of network traffic towards encrypted payloads.

- UNB IDS datasets – these datasets are well documented and have been well maintained, and include both raw events and significant flow metadata, and a broad spectrum of well-labeled threat and anomaly events [49, 88]. Despite being synthetic this is perhaps the most useful dataset available today for IDS research. Note that the dataset is published as part of a much broader archive of threat event data, and security researchers are encouraged to consider all of the datasets available.
- MAWI and CAIDA datasets [23,44] – these are well maintained and extensive archives of live network backbone traffic (including threats and anomalies), with strong provenance and supporting information. Since these datasets deal primarily with Internet backbone data, this is a narrow context for IDS research; albeit an important one. Researchers should understand that event types and distributions within these archives are likely to differ markedly from enterprise, cloud, and other infrastructure types.

The third dataset we highlight despite its age, due to rarity, since this holds SCADA event data that is particularly under-represented in the IDS corpus. This perhaps illustrates the informed compromises that may occasionally have to be made by researchers in this field.

- ICS 2014 - offers unique insight into typical SCADA network threats. Whilst this dataset is quite old, to the best of our knowledge there is no other dataset available that covers this specific domain [89].

We do not discourage the use of other datasets for comparative or teaching purposes, and we acknowledge the debt owed by the IDS research community to the authors of earlier datasets; it is however an unfortunate fact that as a general rule IDS dataset becomes quickly obsolete unless they are rigorously designed and regularly updated.

5 FURTHER ANALYSIS

The author has performed analysis of a number of these datasets. Analysis was performed using a specially created tool, called HIVE [16]. HIVE translates event information from various datasets into rich data structures called Correlated Flow Format (CFF). These flow summaries can be constructed directly from original event files (e.g. PCAP format), or by translating any flow summaries provided with a dataset. CFF flow data can be used to perform statistical analytics across a dataset, and HIVE integrates support for directional analysis of event flows, asymmetry, payload entropy, burst characteristics and flow timing. HIVE places these CFF containers on a queue (called the Flowbus), which enables queue subscribers (e.g. IDS detectors, production rule engines, and various feature mapping and amplification tools) to be orchestrated in various topologies (stacked, chained in series, or run as parallel ensembles for example).

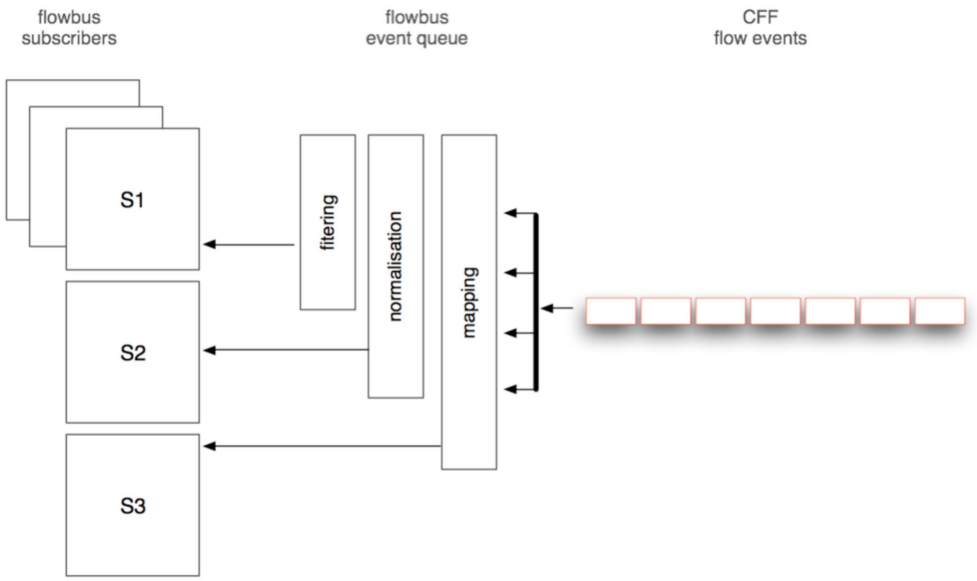


Figure 5: The HIVE Flowbus subscription model, with the possibilities for cooperative, parallel or serialised learning. CFF flow data can be accessed transformed in a number of ways to suit the subscribers, including transformation by subscribers on behalf of downstream clients. Flow events can be generated directly using the HIVE event simulator, or imported from public datasets or network traces.

By using the HIVE analyser on various datasets we can examine normal and threat event distributions, by protocol types, over different time intervals. We can quickly establish whether the dataset appears representative of current trends, and identify low-level flow features that may be promising candidates for inputs into detection and learning models. For example:

- Flow Asymmetry: symmetry has been modeled at both a packet and byte level, with metrics produced in the range [0..1] to denote if the flow is dominated by source or destination. Asymmetry differences between 'normal' and 'attack' traffic can be a good indicator of anomalous behaviour (for example in the case of a SYNflood DDoS attack, only connection requests are normally sent to the target).
- Payload Entropy: since flows or raw trace files contain payload information, an implementation of entropy is included to determine the randomness of payload information (by direction) including Base64 and UTF encodings. Assuming 8-bit character strings, values in the range [0..8] are produced, with 8 indicating maximum entropy. This value can also be normalised in the range [0..1] - for use in neural networks for example. Entropy is a strong indicator of encrypted data, and may be used to identify covert channels or anomalous session use.

5.1.1 Example analysis on ISCXIDS2012

HIVE analysis on the UNB ISCXIDS2012 dataset identified 103.9 million packets, comprising 2.1 million flows, with aggregate anomaly rates of approximately 10.2%. Figure 5 illustrates the 'top-ten' breakdown of normal and threat events, by percentage, illustrating that the dataset is dominated by HTTP web traffic. This is clearly not representative of modern networks - where typically there is higher proportion of HTTPS. We identified 79 discrete threat types (only the top ten are shown in Figure 5), and note that the relative percentage of some of these attacks are as low as 0.00003% - illustrating one of key challenges in designing effective intrusion detection techniques - the signal to noise ratio for anomalous and attack traffic is typically very low.

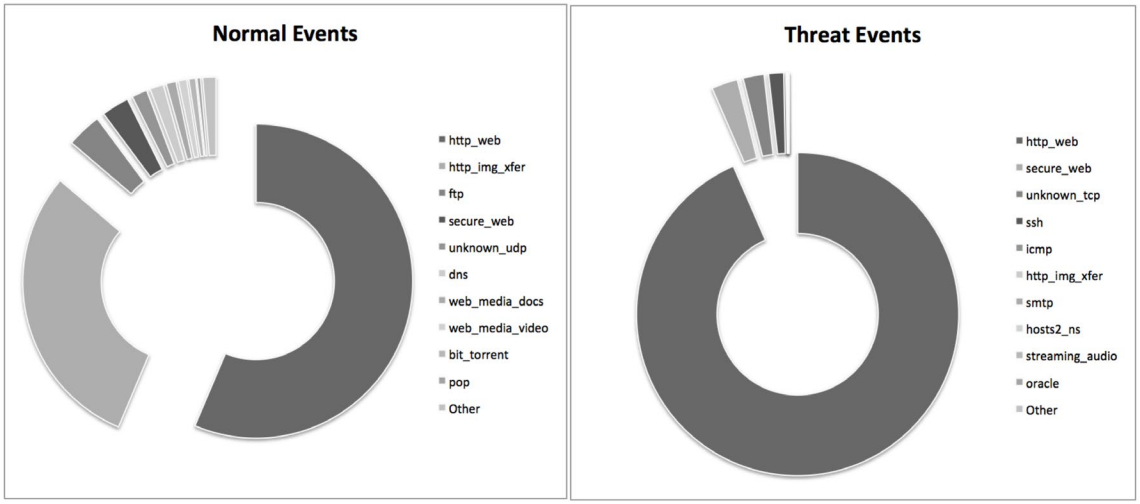


Figure 6: Partial analytics summary report from HIVE, illustrating a number of flow properties calculated from the event data. Note that a full report includes a breakdown of all normal and attack flows, enabling direct comparison of flow characteristics). Additional reports product per-flow analysis across the complete capture period, together with additional metadata.

The UNB ISCX 2012 dataset records flows in XML format, and was found not to record payload consistently for all flows (a small number of flows have no payload recorded), Notably the later UNB datasets move to a more compact CSV summary format, and therefore no payload information is available at that level. This illustrates one of the key advantages of HIVE, in that these flows can be reconstructed using the original PCAP files, and therefore a uniform analysis of entropy (for example) can be performed over all datasets. We intend to extend HIVE to support a number of machine learning libraries, as well as integrating a number of common event formats. We also intend to enable HIVE to provide summary reports on dataset quality and realism, based on some of the best-practice features identified in Figure 4.

A particular feature to note is the trend over time towards encrypted payloads. Older datasets typically exhibit very little encrypted event data; whilst more modern well-maintained datasets (such as those from UNB) will illustrate this shift in traffic composition – especially noticeable in the shift from HTTP to HTTPS web traffic. As a consequence, detection techniques based on flow summary records may be largely unaffected (for example if the techniques rely on flow symmetry, timing, payload size, entropy etc.). However those techniques employing deeper payload analysis, or feature extraction from payload data, may be adversely impacted by the limited availability of features. It may be possible to dynamically unencrypt live payloads with techniques such as SSL interception in some live deployments). It may also be feasible to parse out useful features from plaintext headers with some encrypted traffic (depending on the implementation).

6 Conclusions

Today we face a serious challenge in detecting and preventing cybersecurity breaches and malware threats, and we continue to fall behind in addressing security challenges for large enterprises [33]. The limited availability of high quality, representative, intrusion datasets is a significant obstacle to progress in this field. In this paper, we find that many widely used datasets are either out of date, or not representative of today's threat landscape; in short the majority of public datasets are not fit for purpose, and the value of analysis performed against them may be of questionable value today. Our observations are older datasets are often generated using flawed methodologies, and that new datasets quickly become rapidly obsolete unless routinely updated; primarily due to a rapidly changing threat landscape and real-world network technology

changes. There are some notable exceptions where excellent quality datasets have been well designed and maintained (such as the UNB ISCX IDS datasets), which we highlight in this paper. It should also be noted that some datasets, though well maintained, represent highly specialised domains (such as the MAWI and CAIDA network backbone datasets).

In many cases it is not obvious that the threat event data has been largely synthesised, or at least partly synthetic, and in some cases the threat and normal event distribution may be highly distorted (especially where generated as part of a cyber range simulation). Such datasets may not be suitable as proxies for real-world domains, such as financial institutions, retail, transport, healthcare for example, and given the existing challenges and risks in de-identification, together with potential regulatory concerns, it seems unlikely that domain-specific data for such institutions will be published for wider study. Sharing of sensitive domain-specific data may be possible within closed communities under strict controls, and where the identity of collaborators can be assured and trusted (for example: law enforcement and government establishments).

Issues with maintenance and the limited inclusion of threat events from live environments suggest that future datasets will continue to be produced largely from synthesised sources, with simulated data models validated against live data, with a trend towards heavily programmable virtualised infrastructure to improve maintainability and flexibility.

We acknowledge that many of the datasets reviewed here represented the 'state of the art' at the time of publication, and were themselves invaluable in improving the field of threat detection. Researchers owe a great debt to the authors of these important milestones in intrusion detection. However the reality is that rapid advances in threat sophistication, together with increased dependency on network systems, mean that the scale of the cybersecurity challenge today requires that we must focus research on more agile and representative data models, routinely updated, and exposing a wide range of features through metadata, in order to maintain pace with adversaries.

7 Further Work

Going forward the authors propose several initiatives that would improve security threat research and promote dataset availability:

- Collaboration: much broader collaboration is required between academia, government and commercial enterprises in the creation of open standardized and well-maintained libraries of representative data. Virtualised cloud infrastructure, cooperatively funded, with controlled access, represents a potentially attractive framework to promote collaboration.
- Dataset Quality Measurement: there is a need for standardised metrics for measuring how realistic such data is with regard to live network and system performance, and appropriate presence and distributions of anomalies and threat patterns across various domains.
- De-Identification: further research is needed into the standardisation of de-identification techniques (such as pseudonymisation and anonymisation) to enable sensitive databases to be 'scrubbed' so that they can be published with lower risk to the originating organisation. Whilst a difficult challenge, further advances here could remove some of the barriers in publishing invaluable real-world attack traffic and logs.
- Simulation: further exploration is required in the utility of simulation in generating new attacks and anomalies – based on, and validated against - representative live datasets. Improvements in simulation techniques are required, as well as better instrumentation and metrics to verify simulation behaviour and output. In practice the combination of real and synthesised data is being increasingly adopted to generate more accurate data models [20].
- Virtualised Intrusion Laboratories: Recent advances in programmable systems and cloud computing make it practical to design, simulate and orchestrate large infrastructure using pre-configured virtual machines (VM) and containers. These environments have the benefits of: simplifying test isolation, for running live malware, the ability to emulate large complex physical infrastructure in software, and the ability to programmatically update and maintain datasets for a range of domains.

- Taxonomy & Labeling: researchers need clearer guidance on dataset deployment context, composition, and generation in order to make informed decisions on how applicable a dataset is for the models being developed. It would be helpful if a standard labeling scheme were adopted to classify intrusion datasets.
- Threat Population Metrics: ideally a set of metrics should be published and maintained, per user domain, to enable researchers to model various benign and threat traffic compositions across different vertical industries. Used together with simulation this could (for example) assist in speeding up bootstrapping of domain specific anomaly detection tools, by providing pre-tuned learning models.

REFERENCES

[1] Duffield, N., Haffner, P., Krishnamurthy, B. and Ringberg, H., 2009, April. Rule-based anomaly detection on IP flows. In *IEEE INFOCOM 2009* (pp. 424-432). IEEE.

[2] Kenyon, T., 2018. Transportation Cyber-Physical Systems Security and Privacy. In *Transportation Cyber-Physical Systems* (pp. 115-151). Elsevier.

[3] Ugarte-Pedrero, X., Graziano, M. and Balzarotti, D., 2019. A Close Look at a Daily Dataset of Malware Samples. *ACM Transactions on Privacy and Security (TOPS)*, *22*(1), p.6.

[4] Regulation, G.D.P., 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)*, *59*(1-88), p.294.

[5] Tavallaee M, Stakhanova N, Ghorbani AA. Toward credible evaluation of anomaly-based intrusion detection methods. Trans Sys Man Cyber Part C 2010;40:516e24.

[6] Sommer R, Paxson V. Outside the closed world: on using machine learning for network intrusion detection. In: Security and privacy, IEEE Symposium on; 2010. p. 305e16.

[7] Symantec. Internet Threat Security Report (ITSR) 2019.

[8] Symantec: Global Internet Security Threat Report Trends for 2008. Available on: http://eval.symantec.com/mktginfo/enterprise/white_papers/b-whitepaper_internet_security_threat_report_xiv_04-2009.en-us.pdf

[9] Symantec: 2015 Internet security threat report. Available on: https://www.symantec.com/security_response/publications/ threatreport.jsp.

[10] Mirkovic, J., and Reiher, P., "A taxonomy of DDoS attack and DDoS defense mechanisms," ACM SIGCOMM Computer Communication Review, 2004.

[11] Barnett, R. J. and Irwin, B., "Towards a taxonomy of network scanning techniques," ser. SAICSIT '08, pp. 1–7.

[12] Plonka, D. and Barford, P., "Network anomaly confirmation, diagnosis and remediation," ser. CCC '09, pp. 128–135.

[13] "Capec", Available on: http://capec.mitre.org/.

[14] Mazel, J., Fontugne, R. and Fukuda, K., 2014, August. A taxonomy of anomalies in backbone network traffic. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2014 International* (pp. 30-36). IEEE.

[15] Grégio, A.R.A., Afonso, V.M., Filho, D.S.F., Geus, P.L.D. and Jino, M., 2015. Toward a taxonomy of malware behaviors. *The Computer Journal*, *58*(10), pp.2758-2777.

[16] Ioblox cybersecurity research space, maintained by the first author. Available on: http://ioblox.net/cs/cs-index.html, April 2019.

[17] Security Onion, open source linux sitribution for intrusion detection, security monitoring and logging. Available at: https://securityonion.net, Nov 2018

[18] Glasser, J. and , B., 2013, May. Bridging the gap: A pragmatic approach to generating insider threat data. In *2013 IEEE Security and Privacy Workshops* (pp. 98-104). IEEE.

[19] Haider, W., Hu, J., Slay, J., Turnbull, B.P. and Xie, Y., 2017. Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling. *Journal of Network and Computer Applications*, *87*, pp.185-192.

[20] Shiravi, A., Shiravi, H., Tavallaee, M. and Ghorbani, A.A., 2012. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *computers & security*, *31*(3), pp.357-374.

[21] Creech, G., and Hu, J. (2013). “Generation of a new IDS test dataset: time to retire the KDD collection,” in Proceedings of the 2013 IEEE Wireless Communications and Networking Conference (WCNC), New York NY, 4487–4492.
[22] Lindauer, B., Glasser, J., Rosen, M., Wallnau, K.C. and ExactData, L., 2014. Generating Test Data for Insider Threat Detectors. *JoWUA*, *5*(2), pp.80-94.
[23] Fontugne, R., Borgnat, P., Abry, P. and Fukuda, K., 2010, November. Mawilab: combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In *Proceedings of the 6th International COnference* (p. 8). ACM.
[24] Vasilomanolakis, E., Cordero, C. G., Milanov, N., andMühlhäuser, M. (2016) “Towards the creation of synthetic, yet realistic, intrusion detection datasets”, Network Operations and Management Symposium (NOMS), IEEE, pp 1209-1214.
[25] Coburg University, CIDDS python event generation scrtipts. Available at: https://github.com/markusring/CIDDS. Nov 2018.
[26] Ring, M., Wunderlich, S., Gruedl, D., Landes, D., Hotho, A.: "Flow-based benchmark data sets for intrusion detection." In: Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS), pp. 361-369. ACPI (2017)
[27] RFC3954 Cisco Systems NetFlow Services Export Version 9, Cisco Systems, 2004
[28] Sperotto, A., Sadre, R., Van Vliet, F., and Pras, A. (2009) "A Labeled Data Set For Flow-based Intrusion Detection", Proc. of the 9th IEEE Int. Workshop on IP Operations and Management (IPOM), Springer, pp 39-50.
[29] Wheelus, C., Khoshgoftaar, T. M., Zuech, R., andNajafabadi, M. M. (2014) “A Session Based Approach for Aggregating Network Traffic Data -The SANTA Dataset”, Proc. of the Int. Conf. on Bioinformatics and Bioengineering (BIBE), pp 369-378
[30] Zuech, R., Khoshgoftaar, T. M., Seliya, N., Najafabadi, M. M., and Kemp, C. (2015) “A New Intrusion Detection Benchmarking System”, Proc. of the 28th Int. Florida Artificial Intelligence Research Society Conference, pp 252-256.
[31] Ring, M., Wunderlich, S., Gruedl, D., Landes, D., Hotho, A.: "Creation of Flow-Based Data Sets for Intrusion Detection”. In: Journal of Information Warfare (JIW), Vol. 16, Issue 4, pp. 40-53, 2017
[32] Wang, K. and Stolfo, S.J., 2004, September. Anomalous payload-based network intrusion detection. In *International Workshop on Recent Advances in Intrusion Detection* (pp. 203-222). Springer, Berlin, Heidelberg.
[33] M-Trends 2019, FireEye Mandiant Services, Special Report, 2019.
[34] Ribaric, S., Ariyaeeinia, A. and Pavesic, N., 2016. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, *47*, pp.131-151.
[35] Sweeney, L., 1997. Guaranteeing anonymity when sharing medical data, the Datafly System. In *Proceedings of the AMIA Annual Fall Symposium* (p. 51). American Medical Informatics Association.
[36] Narayanan, A. and Shmatikov, V., 2006. How to break anonymity of the Netflix prize dataset. arXiv preprint cs/0610105.
[37] Sweeney, L., 2013. Matching known patients to health records in Washington State data. Available at SSRN 2289850.
[38] Dinur, I. and Nissim, K., 2003, June. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 202-210). ACM.
[39] Dwork, C. and Roth, A., 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), pp.211-407.
[40] Dwork, C., McSherry, F., Nissim, K., and Smith, A., Calibrating noise to sensitivity in private data analysis. In TCC, pages 265–284, 2006.
[41] Dwork. C., Differential privacy. In ICALP, pages 1–12, 2006.
[42] Ben-Salem, M. RUU dataset: Available at: http://www1.cs.columbia.edu/ids/RUU/data/
[43] Schonlau, M., DuMouchel, W., Ju, W.H., Karr, A.F., Theus, M. and Vardi, Y., 2001. Computer intrusion: Detecting masquerades. Statistical science, pp.58-74.
[44] Center for Applied Internet Data Analysis (CAIDA), dataset sources. Available on: https://www.caida.org/data, March 2019
[45] Proebstel, E. P. (2008). Characterizing and Improving Distributed Network-based Intrusion Detection Systems (NIDS): Timestamp Synchronization and Sampled Traffic. (Davis, CA: University of California DAVIS).

[46] CAIDA (2002). CAIDA data set OC48 Link A. Available at: https://www.caida.org/data/passive/passive oc48 dataset.xml
[47] CAIDA (2007). CAIDA DDoS Attack Dataset. Available at: https://www.caida.org/data/passive/ddos-20070804 dataset.xml
[48] CAIDA (2016). CAIDA Anonymized Internet Traces 2016 Dataset, Available at: https://www.caida.org/data/passive/passive 2016 dataset.xml
[49] Univerisity of New Brunswick Intrusion, Malware and DDoS datasets. Available on: https://www.unb.ca/cic/datasets/.
[50] Alexander D Kent. Cyber security data sources for dynamic network research. Dynamic Networks and Cyber-Security, 1:37, 2016.
[51] Alexander D. Kent. User-computer authentication associations in time. Los Alamos National Laboratory, 2014.
[52] Turcotte, M.J., Kent, A.D. and Hash, C., 2017. Unified host and network data set. *ArXiv e-prints*.
[53] Xie, M., and Hu, J. (2013). "Image and Signal Processing (CISP), 2013 6th International Congress on," in Proceedings of the Evaluating host-based anomaly detection systems: A preliminary analysis of ADFALD, Vol. 03 (Berlin: Springer), 1711–1716.
[54] Xie, M., Hu, J., and Slay, J. (2014). "Evaluating host-based anomaly detection systems: application of the one-class SVM algorithm to ADFA-LD," in Proceedings of the 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, 978–982.
[55] CIDD cloud masqerade dataset. Available on: http://www.di.unipi.it/~hkholidy/projects/cidd/download.html, March 2019.
[56]Greenberg, S: Using unix: Collected traces of 168 users. Report 88/333/45, Univ. of Calgary, 1988.
[57] T.Lane , C.E. Brodley,: An application of machine learning to anomaly detection. In Proc. of the 20th National Information Systems Security Conference. (1997) 366-380
[58] U Mass Trace Repository (2011). Optimistic TCP ACKing, University of Massachusetts Amherst, Availabel at: http://traces.cs.umass.edu
[59] Nehinbe, J. O. (2011). "A critical evaluation of datasets for investigating IDSs and IPSs researches," in Proceedings of the IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS), 92–97, New York, NY.
[60] Prusty, S., Levine, B. N., and Liberatore, M. (2011). "Forensic investigation of the oneswarm anonymous filesharing system," in Proceedings of the ACM Conference on Computer and Communications Security (CCS) (New York, NY: ACM).
[61] García, S., Grill, M., Stiborek, J., & Zunino, A. (2014) "An Empirical Comparison of Botnet Detection Methods", Computers & Security, Vol. 45, pp 100-123.
[62] CSIC HTTP Dataset 2010. Available on: http://www.isi.csic.es/dataset//.
[63] A. Perez-Villegas, C. Torrano-Gimenez, G. Alvarez. Applying Markov Chains to Web IntrusionDetection. In Proc. of Reunión Española sobre Criptología y Seguridad de la Información (RECSI 2010), pp. 361-366. Publicaciones urv. Tarragona (España), 7-10 Septiembre (2010).
[64] C. Torrano-Gimenez, A. Perez-Villegas, G. Alvarez. An anomaly-based approach for intrusion detection in web traffic. Journal of Information Assurance and Security, vol. 5, issue 4, pp. 446-454. ISSN 1554-1010 (2010).
[65] C. Torrano-Gimenez, A. Perez-Villegas, G. Alvarez, A Self-Learning Anomaly-Based Web Application Firewall. In Proc. of 2nd International Workshop in Computational Intelligence in Security for Information Systems (CISIS 09). Advances in Intelligent and Soft Computing, vol. 63, pp. 85-92, Springer-Verlag. A. Herrero, P. Gastaldo, R. Zunino, E. Corchado, editores. Burgos (España), 23-26 Septiembre (2009).
[66] C. Torrano-Gimenez, A. Perez-Villegas, G. Alvarez, An Anomaly-based Web Application Firewall. In Proc. of International Conference on Security and Cryptography (SECRYPT 2009), pp. 23-28. INSTICC Press. E. Fernández-Medina, M. Malek, J. Hernando, editores. Milán (Italia), 7-10 Julio (2009).
[67] H. Nguyen, C. Torrano-Gimenez, G. Álvarez, S. Petrovic, K. Franke, Application of the Generic Feature Selection Measure in Detection of Web Attacks. In Proc. of International Workshop in Computational Intelligence in Security for Information Systems (CISIS 11), LNCS 6694, pp. 25–32. Editor Á. Herrero and E. Corchado, Springer-Verlag. Torremolinos, Málaga (España), Junio (2011).
[68] C. Torrano-Gimenez, H. Nguyen, G. Álvarez, S. Petrovic, K. Franke, Applying Feature Selection to Payload-Based Web Application Firewalls. In Proc. of International Workshop on Security and

Communication Networks (IWSCN 11), pp. 75-81. Editor Patric Bours. Gjøvic (Noruega). ISBN: 978-82-91313-67-2. 18-20 Mayo (2011).
[69] Sangster, B., O'Connor, T.J., Cook, T., Fanelli, R., Dean, E., Morrell, C. and Conti, G.J., 2009, August. Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets. In *CSET*.
[70] Song, J., Takakura, H., Okabe, Y., Eto, M., and Inoue, D., and Nakao, K. (2011). "Statistical analysis of honeypot data and building of kyoto 2006+ dataset for NIDS evaluation," in Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (New York, NY: ACM), 29–36.
[71] Sato, M., Yamaki, H., and Takakura, H. (2012). "Unknown attacks detection using feature extraction from anomaly-based ids alerts," inProceedings of the IEEE/IPSJ 12th International Symposium on Applications and the Internet (SAINT), Piscataway, NJ, 273–277.
[72] Chitrakar, R., and Huang, C. (2012). "Anomaly based intrusion detection using hybrid learning approach of combining k-medoids clustering and naive bayes classification," in Proceedings of the 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), Piscataway, NJ, 1–5.
[73] Sperotto, A., Sadre, R., Van Vliet, F. and Pras, A., 2009, October. A labeled data set for flow-based intrusion detection. In *International Workshop on IP Operations and Management*(pp. 39-50). Springer, Berlin, Heidelberg.
[74] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.
[75] "Nsl-kdd data set for network-based intrusion detection systems." Available on: https://www.unb.ca/cic/datasets/index.html, March 2009.
[76] McHugh, J., 2000. Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Transactions on Information and System Security (TISSEC)*, *3*(4), pp.262-294.
[77] Nechaev, B., Allman, M., Paxson, V. and Gurtov, A., 2004. Lawrence berkeley national laboratory (lbnl)/icsi enterprise tracing project. *Berkeley, CA: LBNL/ICSI*.
[78] The UCLA CSD 2001 packet trace dataset. Available on: https://lasr.cs.ucla.edu/ddos/traces/
[79] DEFCON 8, 10 and 11, The Shmoo Grouphttp: //cctf.shmoo.com, 2000.
[80] Nehinbe, J. O., and Weerasinghe, D. (2010). "A Simple Method for Improving Intrusion Detections in Corporate Networks," in Information Security and Digital Forensics First International Conference ISDF (Berlin: Springer), 111–122.
[81] KDD99. Third International Knowledge Discovery and Data Mining Tools Competition, May 2002, Available from: http://kdd.ics.uci.edu/databases/kddcup99.kddcup99.html.
[82] DARPA Intrusion Detection Evaluation data sets (1998, 1999, 2000). Lincoln Laboratory web page (MIT). http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/.
[83] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. Kendall, D. McClung, D. Webber, S. Webster, D. Wyschograd, R. Cunninghan, and M. Zissman. Evaluating Intrusion Detection Systems: The 1998 DARPA off-line intrusion detection evaluation. In Proc. of DARPA Information Survivability Conference and Exposition (DISCEX00), Hilton Head, South Carolina, January 25-27. IEEE Computer Society Press, Los Alamitos, CA, 1226 (2000).
[84] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba and K. Das. The 1999 DARPA Off-Line Intrusion Detection Evaluation. In Proc. Recent Advances in Intrusion Detection (RAID2000). H. Debar, L. Me, and S. F. Wu, Eds. Springer-Verlag, New York, NY, 162182 (2000).
[85] Brown, C., Cowperthwaite, A., Hijazi, A., Somayaji, A. (2009). "Analysis of the 1999 DARPA/lincoln laboratory IDS evaluation data with NetADHICT," in Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications, Piscataway, NJ, 1–7.
[86] M. Mahoney and P. Chan, "An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection," *LEC- TURE NOTES IN COMPUTER SCIENCE*, pp. 220–238, 2003.
[87] CTU University, Czech Republic, dataset: Available on: https://www.stratosphereips.org/datasets-ctu13/
[88] CSE-CIC-IDS2018 on AWS. A collaborative project between the Communications Security Establishment (CSE) & the Canadian Institute for Cybersecurity (CIC) on: https://www.unb.ca/cic/datasets/ids-2018.html

[89] Industrial Control Systems Cyber Attack Datasets, on: https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets.
[90] The Royal Academy of Engineering. *Smart Infrastructure: The Future*; The Royal Academy of Engineering: London, UK, 2012; pp. 16–17. ISBN 1-903496-79-9.
[91] The Aegean WiFi Intrusion Dataset (AWID), an 802.11 intrusion dataset from the University of the Aegean, available since 2015. Available on: http://icsdweb.aegean.gr/awid/index.html. January 2020.
[92] Kolias, C., Kambourakis, G., Stavrou, A. and Gritzalis, S., 2015. Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset. *IEEE Communications Surveys & Tutorials*, *18*(1), pp.184-208.
[93] Garcia-Font, V., Garrigues, C. and Rifà-Pous, H., 2018. Difficulties and challenges of anomaly detection in smart cities: A laboratory analysis. *Sensors*, *18*(10), p.3198.
[94] NYUAD. Smart City Testbed NYUAD. Available on: http://sites.nyuad.nyu.edu/ccs-ad/about/research-areas-2/research-labs-groups/smart-city-testbed/. January 2020).
[95] UNSW-NB15 datasets from the University of New South Wales. Available on: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
[96] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.
[97] ADFA dataset resource links at the University of New South Wales. Available on: https://research.unsw.edu.au/people/professor-jiankun-hu
[98] Moustafa, N., Turnbull, B. and Choo, K.K.R., 2018. An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of internet of things. *IEEE Internet of Things Journal*, *6*(3), pp.4815-4830
[99] Internet Engineering Task Force, 2013a. *RFC-7011 Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information.* Available on: https://tools.ietf.org/html/rfc7011
[100] Khraisat, A., Gondal, I., Vamplew, P. and Kamruzzaman, J., 2019. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, *2*(1), p.20.
[101] Hindy, H., Brosset, D., Bayne, E., Seeam, A., Tachtatzis, C., Atkinson, R. and Bellekens, X., 2020. A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems. *IEEE Access*.